MASTER'S DEGREE THESIS

MSc in Statistics and Operations Research

Title: Permutation multivariate analysis of variance on real data and simulations to evaluate for robustness against dispersion and unbalancedness.

Author: Lucas Tomas Noce

Advisor: Jordi Ocaña Rebull

Department: Statistics

Academic year: 2013-2014

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

Master's degree thesis: Permutation multivariate analysis of variance on real data and simulations to evaluate for robustness against dispersion and unbalancedness

Lucas Noce

October 5, 2014

Contents

List of Figures

List of Tables

1 Introduction

This work was motivated by the analyses of some ecological data, collected between 2011 and 2012 near Puerto Madryn in Patagonia, Argentina. Researchers wanted to know whether the presence of seabird nesting colonies, which was the factor of interest, had an effect on some nutrients and chlorophyll in the coastal area. Other than the month of observation, data was collected also according to several locations distributed near Puerto Madryn. Data presented some signs of heteroscedasticity. Furthermore, the number of replicates across the different combinations of the factors' levels were unequal and some combinations were missing. Both patterns might induce different data dispersions across factors' levels.

Such unbalanced data structure inspired us to perform simulation studies with the aim of elucidating, empirically, to what extent distance-based permutational methods are robust when data have different dispersions across factors' levels. Dispersion was induced by means of both different variances across factors' levels and by generating replicates of unequal sample sizes and even with missing data.

The analyses, described in Section 4, were performed by means of a distance-based permutational ANOVA method, the so-called PERMANOVA. Such procedure is based on the concept of partitioning a distance matrix according to the factors of interest, bringing reliable results as in traditional MANOVA but allowing for several data structures, far from normally distributed.

The present study consisted of two different parts. Part I (Sections 2 and 3) covered an introduction to the theoretical aspects of PERMANOVA and a simulation study in which 432 scenarios were empirically analysed in order to take account of the data structure observed, presented in Part II. Given the high number of scenarios, in order to identify patterns or extreme cases, we employed, mainly, data visualization tools.

In Part II (Sections 4 and 5), the above-mentioned real data was described and analysed by means of a PERMANOVA test, with the aim of answering the main questions of the researchers. Conclusions, reported in Section 5, took account of our results on robustness derived from the simulation study, with the aim of evaluating the found P-values in the light of the permormance manifested during the simulations. Detailed tables about the simulations' results were included in the Appendix.

Figure 1: Patagonia, Atlantic Ocean.

Part I
Analysis of PERMANOVA by simulations

2 The general approach

A procedure of permutation tests based on distances, from now on PTBD, employs the following elements:

1. A distance measure related to the observations;
2. A statistic, function of the distance measure;
3. A distribution of the statistic built through permutation of data;
4. A decision rule based on the so called P-value, which is computed by counting permutations.

Since the field of application is the analysis of variance, some other important elements are considered:

1. The variance partitioning;
2. The expected means;
3. The relation bewteen permutations and experimental design matrix.

The following sections cover all these elements.

2.1 Distances, dissimilarities and data

2.1.1 Distance as association

2.1.1.1 Distance as association In the present section and the followings, the term *association* is intended as a term which is measured in order to quantify any resemblance or dissimilarity between objects, following Legendre, P. and Legendre, L. (1998). A distance is a measure of association between N given objects $\omega_1, ..., \omega_N$ -typically a row in a data matrix- which is computed through the p variables $y_{1i}, ..., y_{pi}$ for $i = 1, ..., N$ that characterize them. For example, y_{1i} denotes the first component or descriptor -the measured variable- of the object -observation- ω_i. In the context of multivariate anaylsis, $p > 2$. Indeed, $y_i' = (y_{1i}, ..., y_{pi})$ denotes, in general, the multiresponse of object ω_i. The multiresponse vector y_i' will be also designated with bold letters $\boldsymbol{y}_i$, as is current in the matrix notation.

An association is intended to be either a difference (dissimilarity) or a resemblance, the nature of which is related to the measure of choice. Since measures of association are usually structured in matrices, we get an association matrix (for example, a distance matrix - an association matrix whose elements are distance coefficients) which is intended to offer a summary of the information in the original data while keeping the strucure. The pursuied structure obeys to the purpose of the problem or hypothesis. So, for example, the euclidean distance is not a reliable measure to compare sites on the bases of species abundancies in ecology, due to the double zero paradox (Legendre, P. and Legendre,L., 1998). As far as the exposition of the first part of this work is concerned, the choice of the distance measure is not of primary importance, since it is related to the purpose of the problem. In a PTBD procedure, one is given a distance matrix, and the procedure itself should guarantee that any distance measure could be used as a starting point of the procedure.

A distance measure, generally speaking, is maximum when two objects are completely different across descriptors and minimum when they are identical.

2.1.1.2 Metrics and semimetrics A distinction relevant in the context of PTBD is among metrics and semimetrics. According to some authors, the distintion is bewteen distances (metric properties) and dissimilarity coefficients (Legendre, P. and Legendre,L., 1998). The distinction is important because some ecological and genetical studies may use, as part of the PTBD procedure, semimetric distances which better conserves the information structure of the original data. In traditional analysis of variance procedures, the use of semimetric measures such as Bray-Curtis, may be not possible due to the mathematical restrictions: in the case of the Euclidean distance, the average of each component of y_i' across the observations withing a group of interest equals its centroid, which is the measure of central location for that group. For the Bray-Curtis distance, a simple average doesn't corresponds to the central location.

Being a, b, c three objects, a distance measure D fulfils the following properties:

1. Minimum 0: if $a = b$ then $D(a, b) = 0$;
2. Positiveness: if $a \neq b$ then $D(a, b) > 0$;
3. Symmetry: $D(a, b) = D(b, a)$;
4. Triangle inequality: $D(a, b) + D(b, c) \geq D(a, c)$.

A metric has 0 value if the two objects are identical and is positive if objects are different. The importance of the last property is that it makes possible the ordering of objects in the eucliedean space: the statement "the sum of distances from a to b and from b to c is greater than the distance from a to c" has indeed sense if the triangle property is accomplished. Semimetrics do not follow property (4) and it means that, unlike the metric, they cannot be used to order points in the Euclidean space: in this case the sum of the distances between a and b and b and c might be smaller than the distance between a and c.

2.1.2 Distance measure

2.1.2.1 Measures There are several distance measures d_{ij} between two objects ω_i and ω_j and each one has sense according to the objects being evaluated.

One of the most frequent used metrics is the Euclidean distance, which might be not proper in the case of species abundance data (Legendre, P. and Legendre,L., 1998). For two objects ω_i and ω_j -two observations in a data matrix measured by p descriptors $y_{1i}, ..., y_{pi}$ and $y_{1j}, ..., y_{pj}$, the euclidean distance is defined as:

$$d_{ij} = \sqrt{\sum_{h=1}^{p} (y_{hi} - y_{hj})^2}$$

A general expression given in Mielke, P.W.Jr. and Berry, K.J (2001) which allows for the euclidean and the squared euclidean is

$$d_{ij} = [\sum_{h=1}^{p} (y_{hi} - y_{hj})^2]^{\frac{v}{2}}$$

When v=1, then d_{ij} is the euclidean distance and when v=2 d_{ij} is the squared euclidean distance, which is non metric and associated with usual parametric tests as the t-test and all ANOVA methods. Indeed Mielke,P.W.Jr. and Berry,K.J (2001) advocate to employ $v = 1$ which results in the Euclidean distance due to robustness considerations and a congruence between the data and analysis spaces: since data space is perceived as an Euclidean space, the choice of $v = 1$ is appealing.

Other distance measures are very frequent in particular data sets. Very common in the study of ecological communities is for example the Bray-Curtis measure of dissimilarity, definded as:

$$d_{ij} = \sqrt{\frac{\sum_{h=1}^{r} |y_{hi} - y_{hj}|}{\sum_{h=1}^{r} (y_{hi} + y_{hj})}}$$

2.1.2.2 Distance matrix Distance measures will normally be represented as square symmetric matrices with diagonal elements equal to zeros.

Given N samples of vector observations $y_{1i}, ..., y_{pi}$ where i is any given sample and p, the number of variables, then all the distances between the N objects can be ordered in the following matrix:

$$\begin{bmatrix} d_{11} & d_{12} & d_{13} & d_{14} & ... & d_{1N} \\ d_{21} & d_{22} & d_{23} & d_{24} & ... & d_{2N} \\ d_{31} & d_{32} & d_{33} & d_{34} & ... & d_{3N} \\ d_{41} & d_{42} & d_{43} & d_{44} & ... & d_{4N} \\ ... & ... & ... & ... & ... & ... \\ d_{N1} & d_{N2} & d_{N3} & d_{N4} & ... & d_{NN} \end{bmatrix} \tag{1}$$

Any element d_{ij} represents the value of a distance coefficient, either Euclidean, Bray-Curtis or any choice according to the object of study. Due to properties of minimum and symmetry, the above distance matrix can be also represented as:

$$\begin{bmatrix} 0 & & & & & \\ d_{21} & 0 & & & & \\ d_{31} & d_{32} & 0 & & & \\ d_{41} & d_{42} & d_{43} & 0 & & \\ ... & ... & ... & ... & ... & \\ d_{N1} & d_{N2} & d_{N3} & ... & d_{N(N-1)} & 0 \end{bmatrix} \tag{2}$$

Actually it becomes a lower or upper triangular matrix. The distance matrix is an important object because the PERMANOVA procedure is based on a partitioning of the above matrix in the same way as observations can be partitioned according to the "within" groups or "between" groups in the classical analysis of variance.

2.2 Significance and permutation tests

2.2.1 Level of significance

2.2.1.1 P-value Permutations, in experimental design, are based on resampling the data under several schemes according to the design. It is similar to the bootstrap procedure but resampling is done without replacement.

For a given observed sample of vector observations (of dimension p),

$\boldsymbol{y} = y'_{11}, ..., y'_{1n_1}, y'_{21}, ..., y'_{2n_2}, ..., y'_{g1}, ..., y'_{gn_g}$

of size $N = \sum_{j=1}^{g} n_j$ where g is the number of groups among which the sample can be splitted according to some given criteria of interest, a given statistic $t(\boldsymbol{y})$ is computed which considers the information of the factor levels. Then, the P-value of $t(\boldsymbol{y})$ is defined by

$$P - value = P_{H_0}[t \geq t(\boldsymbol{y})] \text{ or } P - value = P_{H_0}[t \leq t(\boldsymbol{y})]$$

depending on the rejection area that is defined by the hypothesis. Here, the suffix H_0 means the null hypothesis of equal distributions among the g groups. If the given statistic t belonged to a given distribution model (parametric view), then $P_{H_0}[t \geq t(\boldsymbol{y})]$ could be computed by reference to the distribution model. Alternatively, let $\boldsymbol{y'}$ be a resampling without replacement of data $\boldsymbol{y}$ and $t(\boldsymbol{y'})$ the above given statistic based on the resampled data. Data $\boldsymbol{y}$ could be resampled P times and $t(\boldsymbol{y'})$ computed accordingly, P times. Then the above P-value could be obtained numerically as:

$$P - value = \frac{\#(t(\boldsymbol{y'}) > t(\boldsymbol{y}))}{P} \text{ or } P - value = \frac{\#(t(\boldsymbol{y'}) < t(\boldsymbol{y}))}{P}$$

depending on whether big or small values of t are considered to be a rejection of the null hypothesis. In this case the suffix H_0 means that resampling is done according to the null hypothesis of random (by chance) arrangement of data according to the design. It is up to the researcher to establish whether the level of significance is significant. Usually is considered significant when it is established a decision rule as:

$$P - value = \frac{\#(t(\boldsymbol{y'}) > t(\boldsymbol{y}))}{P} < \alpha \text{ or } P - value = \frac{\#(t(\boldsymbol{y'}) < t(\boldsymbol{y}))}{P} < \alpha$$

In a PTBD procedure the information associated to the original data is reduced to a test statistic F^* whose distribution is not known in advance since no distributional assumptions were made for the samples. The convenience of refering to the statistic as F^* is by analogy to the F-ratio statistic frequent in analysis of variance. Each time the data is permuted, a statistic F can be computed, so the set of all these F (including F^*) constitutes the distribution of the statistic under the null hypothesis H_0, which is the reference set for determining the significance (Edgington, E.S., 1995). Choosing a statistic is a way to measure a pattern in the data that the experimenter wants to test. The null hypothesis is that the observed pattern occured only by chance so the observed F^* is one of the many that could have appeared by chance. How *typical* under the assumption of true H_0 F^* is, conducts to the experimenter decision.

2.2.2 Assumptions

2.2.2.1 Why permuting The most important advantage of permuting as a way to test for a hypothesis is that there is no need of making parametric assumptions regarding

the distribution of data. This might be important in the context of many data sets, for example in the case of species in an ecological data set. The normal multivariate assumption, at the base of the parametric MANOVA, might not be meet where species are the variables due, for example, to high skewness or the presence of many zeros (Legendre, P. and Legendre,L., 1998).
Other usual assumptions at the base of parametric methods, such as homoscedasticity, regular exponential familiy or random sampling might not be meet as well. We'll see, nevertheless, that the PERMANOVA procedure, even though is based on permutations, is sensitive to the failure of homoscedasticity.

2.2.2.2 Exchangeability The distribution of the F statistic is obtained through resampling of data *under the null hypotesis*. In a multivariate approach this means that, being

$$\begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \cdots & \cdots & \cdots & \\ y_{N1} & y_{N2} & \cdots & y_{Np} \end{bmatrix} \tag{3}$$

the $N \times p$ matrix where N is the number of samples and p the number of variables, multivariate rows would be exchangeable. This is important because if all y_{ij} values were unrestrictedly resampled, then the dissimilarity structure of the data would be broken. Each row in the above matrix is labeled as belonging to a given group or factor level or combination of groups - to the a group of factor A or to the ab crossed classification of factors A and B, etc. The idea behind permuting the entire rows *under the null hypotesis* is to consider the labels of each row (as belonging to the groups) as random. So considering the rows as echangeable means that the labels can be shuffled. The full set of values in the distance matrix do not change, but their location according to the group structure may differ from one permutation to another or to the original data.

The only assumption that is required to be met is that observations, wich are rows of the original multiresponse data matrix, are exchangeable under the null hypothesis (Anderson, M. J., 2001). If this were the case, a formal consequence is that distributions across factors' levels would be the same, not only in their location but also in their dispersions. Exchangeability, when more than one factor is believed to have an effect, is an assumption regarding the factors' combinations of interest. Being A and B two any given factors (it can be extended to more than two), studying through permutations the effect of, say, factor A, means that exchangeability must be assumed across A-levels but avoiding shuffling the labels across B-levels, becuase it was not assumed that B might have indeed an effect.

2.2.2.3 Dispersions Assuming exchangeability is equivalent to assuming independent observations and similar distributions, which means similar dispersion patterns among groups. Figure 2 shows central locations and dispersions of a single set of simulated multivariate observations whose rows are labeled as belonging to two levels of a single factor. Samples were simulated under equal locations but different dispersions in each of the two

levels. The plot shows the two principal coordinates after multidimensional scaling.

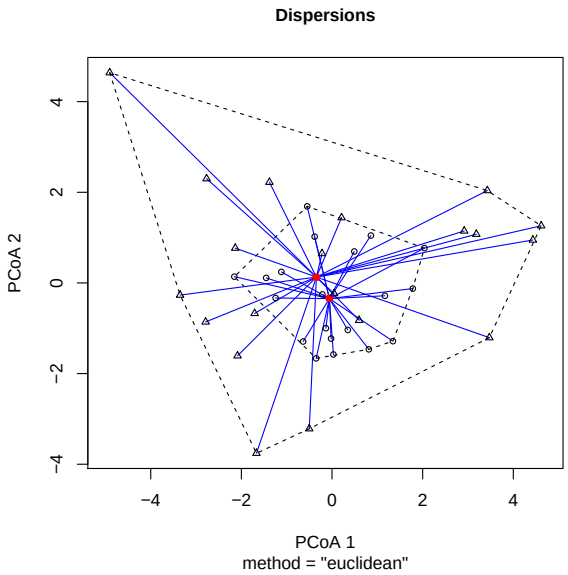

Figure 2: Central locations and dispersions of a set of simulated multivariate variates splited into two groups with different variance-covariance matrices and ploted after mutidimensional scaling.

2.2.3 Types of permutations

2.2.3.1 Monte Carlo approach to permutations Permuting requires a complete enumeration of the resamples in order to build the distribution of reference. Given n replicates and 2 groups there are $n!/r!(n-r)!$ ways to divide the observations into the groups of size r and $n-r$. For 20 observations and two groups of 10 this would be 184765, which is reasonable, but that number grows fast as the number of replicates increases -for 1 replicate more, the number of ways to allocate 21 observations into two groups of 10 and 11 elements is 352716. So one approach is to consider a sample of the total number of possible combinations or a sample of the possible permuted statistics. This is the Monte Carlo approach, which requires randomly sampling from the whole set of permutations considerig that each permuted statistic has the same probability of being chosen.

The Monte Carlo approach is a reasonable choice also when there are very few permutated statistics. For example, if due to the number of observations and size subgroups the number of exact permutations is just 5, then the smallest P-value that can be obtained is 0.2, so it seems a situation in which there is no enough information to build a permutation distribution. Anderson *et al.* (2003) showed that the numerator of the F statistic has, under permutation, an asymptotic distribution which is a linear combination of chi-sqaured random variables. So these random variates can be simulated using Monte Carlo sampling. This case is different from a pure permutation since the distribution is built from a parametric model.

2.2.3.2 Exact test Let consider a given statistic $\boldsymbol{t}$, built on the data, to test a given hypothesis regarding the effect of a factor of interest on some observed variable of interest. And let the level of significance, defined as α, be the maximum probability accepted of commiting the so called type I error, i.e. rejecting the null hypothesis H_0 of absence of effect. In the parametric ANOVA, if H_0 is true, the probability of rejecting H_0 when it is true is exactly equal to α, because the distribution of $\boldsymbol{t}$ is known. Since $\boldsymbol{t}$ is known, the probability of commiting the type I error is always α for every sample size n. When ANOVA is performed through permutations, the test is exact whenever it is possible to enumerate all possible permutations, given a sample size: a sample size n that allows to enumerate all permutations, allows to build the distribution of $\boldsymbol{t}$. When n is so that it is impossible to enumerate all possible permutations, then the distribution of $\boldsymbol{t}$ is built upon a random sample m of total permutations. Then it is possible to approximate the true distribution by increasing m.

In multifactorial ANOVA designs, some permutation tests are exact while others are only approximate and asymptotically approach the exact test. Whether exact or not depends on the design and on the factors that are of interest, so on the given hypothesis. As explained in the next section, if the factor of interest or permutable unit allows for a complete reordering, then an exact text can be built. In other situations, it can happen that, after restricting for those factors that are not of interest, not enough reorderings are available for the permutable unit, limiting the construction of the distribution of $\boldsymbol{t}$. In these cases, only aproximate tests must be constructed.

2.2.3.3 Permutable unit We have seen that resampling of entire rows in a $N \times p$ data matrix is a way to build a reference distribution for a statistic based on that data. In mutifactorial design the concept of permutable unit is very important, since not always the available units to permute allow for an exact test. A permutable unit is that part of a data matrix that will be resampled in order to build the statistic distribution of reference. Anderson ,M.J. *et al.* (2003) have considered different cases as a function of the underlying design in multifactorial situations, including nested, pure random and mixed models.
Some important examples in the case of experimental design are:

1. Unrestricted permutation of raw data. It means complete resampling of raw data independently of the labels which attribute a given observation to a given level of a factor. In the case of a multivariate data matrix, this mean complete resampling of rows only, as explained above.

2. Restricted permutation. In this case the observations that are permuted are those within the levels of another factor whose effect is isolated. Examples are:

   - *2-way crossed fixed design* In the terminology of the sums of squares, it means an exchange of variability between factor A and residuals R (i.e. between SS_A SS_A and SS_R, where the notation SS stands for the "sum of squares" that are the frequent way of reporting results in the analysis of variance) by leaving unchanged the variability attributed to factor B and the total variability (SS_A

and SS_T remain unchanged). If A represents rows and B columns, then units will permute vertically in a way that there is no change of information between the columns.

- *Interaction in a 2-way crossed fixed design* In this case factors A and B would be left unchanged so the permutable units are those withing each *ab* cell. But it is clear that resampling them, would leave unchanged all the quantities involved in the sums of squares partitioning, so no distribution of reference can be built. This leads to other types of permutations such as unrestricted or permutation of residuals.
- *2-way crossed mixed* In order to test the fixed effect A, each sub cell of the levels of A given the levels of B would be permuted. In this case are permuted entire cells vertically, being B the columns.

Other types of permutations are:

1. Permutation of residuals under a reduced model. It is an approximate (not exact) test and consists of isolating the term of interest by fitting the rest of terms and then permuting the residuals unrestrictevily according to the factor's levels of interest. This is indeed a way to test for interaction. Anderson, M.C. *et al.* (2003) have tested its properties for a 2-way crossed fixed design finding that it approaches asymptotically to an exact test because in computing the residuals the fixed effects are excluded by substracting the means. Permuted units are r_{ij} where

   $$r_{ij} = y_{ijk} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}$$

   where $\bar{y}_{i..}$, $\bar{y}_{.j.}$ and $\bar{y}_{...}$ are A's group mean, B's group mean and overall mean respectively. Permuting residuals under a reduced model is the method used by Mielke,P.W.Jr. and Berry,K.J (2001) in applying the MRPP (Multiresponse Permutation Procedure) to multiple regression as modeled for analysis of variance cases. In order to increase the nonparametric character, Mielke, P.W.Jr. *et al.* do a partition of the residuals obtained from a least sum of absolute deviations. It seesm that this procedure assures high resistance to the effect of outliers.

2. Permutation of residuals under a full model. It consists on computing the residuals substracting from each replicate the average corresponding to its particular cell and permute those.

For an exhaustive anaysis of the effects of each type of permutation testing procedure in terms of exactness and power for multifactorial designs, see Anderson ,M.J. *et al.* (2003).

2.3 PERMANOVA

2.3.1 What is PERMANOVA

2.3.1.1 What is PERMANOVA PERMANOVA stands for *Permutation multivariate analysis of variance* and is a testing procedure to detect differences in locations among

groups which rely on dissimilarity measures among samples as the main input information (Anderson, M.C. 2001). Since a distance is a measure of dissimilarity between objets, consider labeling the lower triangular distance matrix

$$\begin{bmatrix} & & & & \\ d_{21} & & & & \\ d_{31} & d_{32} & & & \\ d_{41} & d_{42} & d_{43} & & \\ ... & ... & ... & ... & \\ d_{N1} & d_{N2} & d_{N3} & ... & d_{N(N-1)} \end{bmatrix} \tag{4}$$

according to each group or combination of groups of treatments or characteristics. For example if d_{ij} would so that both samples $y_i^{'}$ and $y_j^{'}$ belonged to the same group of treatment, then d_{ij} would be clasified as a measure of dissimilarity "within" the group. If not, then would be classified as "between" the groups. Based on this information a statistic is built to test for differences among groups. The testing procedure consists of permuting the statistic as explained above.

2.3.1.2 What is it for Classical applications of PERMANOVA are any kind of design in the context of the analysis of variance. As we will see, a partioning of variance is also possible from the distance matrix. The testing procedure allows for any distance measure, either metric or semimetric and this makes of PERMANOVA a very flexible method when it comes to deal with data structures that are not normally distributed. It is also a very reasonable approach to deal with the dimensionality problem, when the number of variables (e.g. species) is greater than the number of cases. Consider for example the case of species abundance. Distributions of this type are very skewed to the right (a species is abundant only in a few observations) and traditional parametric MANOVA relies on the assumption of normally distributed data.

2.3.2 How does it work

2.3.2.1 Geometric approach to MANOVA Consider Figure 3. It is a sample of points from a bivariate normal distribution with clearly two patterns. Each cloud has a centroid and it is possible to compute the distance from each point of each cloud to their respectives centroids. Also it is possible to compute the distance from all points to the overall mean or centroid.

Having this information at hand, it is possible to compute:

1. SS_T: the sum of squared distances from the bivariate points to the overall centroid;
2. SS_W: the sum of squared distances from the samples to their own centroid;
3. SS_E: the sum of squared distances from the group centroids to the overall mean.

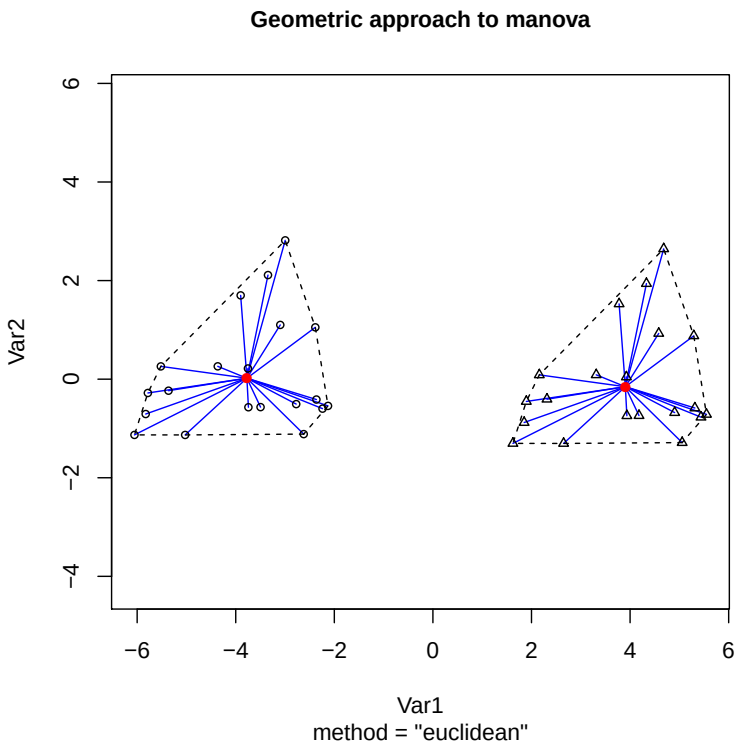

Figure 3: Geometric approach to MANOVA.

For such a partition holds that $SS_T = SS_E + SS_W$. Eventhough it is not the purpose of this work to compare with parametric MANOVA, notice the difference between the geometric approach and the traditional MANOVA. Considering the case of just one factor with g levels and being $\boldsymbol{y}$ a bivariate vector of variables, for the geometric approach the sum of squares SS_W

$$\sum_{i=1}^{g} \sum_{j=1}^{n_i} (\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_{i.})'(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_{i.})$$

is actually a scalar based on a inner product, while for parametric MANOVA it is a matrix based on an outer product

$$\sum_{i=1}^{g} \sum_{j=1}^{n_i} (\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_{i.})(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_{i.})'$$

2.3.2.2 Huygens' theorem The above partitioning holds in the euclidean space since it implies a direct relationship between distance and centroids based on sample means. But as was already mentioned it is not always the case for some distance measures. In this cases it is possible to rely on the Huygens' theorem, according to which the sum of squared distances from each point to their centroid equals the average sum of squared interdistances:

$$\sum_{j=1}^{N} (y_j - \bar{y})^2 = \frac{1}{N} \sum_{1 \leq i < j \leq N}^{N} (y_i - y_j)^2$$

Note that the first sum needs the calculation of a centroid (difficult to achieve from the raw data when distances such as Bray-Curtis are considered) while the second needs only the coordinates.

2.3.2.3 Total sums of squares based on distances Huygen's theorem allows to express the sums of squares differences in relation to a centroid in terms of the average of the squared interdistances. Being d_{ij} a given distance measure based on the data and collected in the lower triangular distance matrix (or upper triangular due to the property of simmetry), then:

$$SS_T = \frac{1}{N} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_{ij}^2$$

This is equivalent to adding up all the squared lower-diagonal elements of

$$\begin{bmatrix} & & & & \\ d_{21} & & & & \\ d_{31} & d_{32} & & & \\ d_{41} & d_{42} & d_{43} & & \\ ... & ... & ... & ... & \\ d_{N1} & d_{N2} & d_{N3} & ... & d_{N(N-1)} \end{bmatrix} \tag{5}$$

2.3.2.4 Structure of the distance matrix $\boldsymbol{D}$ Given a distance matrix $\boldsymbol{D}$ it is possible to code each entry in relation to the experimental design, by partitioning $\boldsymbol{D}$ according to the groups of interest (levels of A in a one-way design or levels of more factors in a multifactorial design). Consider a given d_{ij} which referes to two elements $y_i^{'}$ and $y_j^{'}$ that both belong to the same group. Then

$$SS_W = \frac{1}{n} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_{ij}^2 \phi_{ij}$$

where n is the length of the subgroup or subgroups when the design is balanced and ϕ_{ij} is an indicator which relates the distance to the group of observation:

$$\phi_{ij} = \begin{cases} 1 & if observations\ y_i' \ and\ y_j' \ belong\ to\ the\ same\ group \\ 0 & otherwise \end{cases}$$

2.3.2.5 Pseudo F-ratio statistic Given that the sum of squares due to the hypothesis or factor, say A is $SS_A = SS_T - SS_W$, a pseudo F-ratio statistic to test for multivariate differences in groups (or for a significant effect of factor A) is:

$$F_A = \frac{SS_A/(g-1)}{SS_T/(N-g)}$$

where $(g-1)$ are the degrees of freedom due to the factor and $(N-g)$ the residual degrees of freedom.

The interpretation of the pseudo F-ratio is the same as in the case of the parametric procedure: as F_A grows, meaning that the variability bewteen groups is not only due to the so called experimental error and any other residual error, then the likelihood of the null hypothesys being true lowers. All the above explained quantities are suitable for a multivariate analysis of variance since the distance measure d_{ij} between two objects i and j can be computed for each of the p descriptors. But in the case that d_{ij} is an euclidean distance -and then it is squared according to the previous quantities SS_T and SS_W- and the response or descriptor is univariate, then F_A is the same as the Fisher's F-ratio. Note that even though in this special case both quantities are the same statistic, the way in which their observed signifance is computed is different, because the parametric F relies on a known distribution model for the raw data (objects sampled from a normal distribution). The distribution of F_A is built through permutations of data, so is not known in advance and it is called *pseudo F-ratio* because is based in any given sum of squares of distances and not just the traditional used in parametric ANOVA, the Euclidean.

2.3.2.6 Other way to build an F-ratio An equivalent version of the F_A statistic is obtained by considering ANOVA or MANOVA as linear models. Given a distance matrix $\boldsymbol{D} = [d_{ij}]$, not necessarily euclidean, the Gower centered matrix $\boldsymbol{G}$ is computed according to:

$$\boldsymbol{G} = (\boldsymbol{I} - \tfrac{1}{N}\mathbf{1}\mathbf{1}')\boldsymbol{A}(\boldsymbol{I} - \tfrac{1}{N}\mathbf{1}\mathbf{1}')$$

where $\boldsymbol{A}$ is related to the distances through $\boldsymbol{A} = [a_{ij}] = [-\frac{1}{2}d_{ij}^2]$, $\boldsymbol{I}$ is a $(N \times N)$ identity matrix and $\mathbf{1}$ is a $(N \times 1)$ vector of one's.

Given a hat matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ associated to the linear model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, then the equivalent pseudo-F statistic based on any distance is:

$$F_d = \frac{tr(\boldsymbol{H}\boldsymbol{G}\boldsymbol{H})/(g-1)}{tr((\boldsymbol{I}-\boldsymbol{H})\boldsymbol{G}(\boldsymbol{I}-\boldsymbol{H}))/(N-g)}$$

where the numerator represents the sum of squares due to the hypothesis and the denominator is a residuals sum of square. The trace in both the numerator and denominator will consider only the sums of squares and not the cross products, since the expression $\boldsymbol{H}\boldsymbol{G}\boldsymbol{H}$ is an outer product and not a scarlar. The above expression is equivalent to

$$F = \frac{tr(\hat{\boldsymbol{Y}}'\hat{\boldsymbol{Y}})/(m-1)}{tr(\boldsymbol{R}'\boldsymbol{R})/(n-m)}$$

where $\hat{\boldsymbol{Y}}$ is the matrix of fitted values, $\boldsymbol{R}$ the matrix of residuals $\boldsymbol{R} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$, n the number of observations and m the number of dummy variables associated with the levels of a given factor. The meaning of the numerator and denominator follows from the fact that, beginning with a linear model, the outer product matrix of sums of squares and cross products $\boldsymbol{Y}'\boldsymbol{Y}$ is partitioned into fitted and residual matrices

$$\boldsymbol{Y}'\boldsymbol{Y} = \hat{\boldsymbol{Y}}'\hat{\boldsymbol{Y}} + \boldsymbol{R}'\boldsymbol{R}$$

and the the trace operator is applied

$$tr(\boldsymbol{Y'Y}) = tr(\hat{\boldsymbol{Y}}'\hat{\boldsymbol{Y}}) + tr(\boldsymbol{R'R})$$

which leads to the elements needed for the F statistic. But the importance of the expression F_s resides in its flexibility to partition a sum of squares based on a distance matrix from the data. And its use is possible also when the only available information is a distance matrix.

2.3.2.7 Permutations The pseudo F-ratio has no known distribution since non assumption on the probabilistic model for the raw data are made. The observed F_A^* is compared to a distribution of reference buil by permutations of exchangeable units as explained in Subsection 2.2. Under the null hypothesis the asociation between observed samples and factor labels is considered pure random in the sense that all labels for a given sample are equally likely. This suggests randomly shuffling factor labels among data and for each rearrangement compute the associated F^π. After repeating this operation enough times, the significance of the observed F_A^* is computed as follows:

$$P - value = \frac{(\#F^\pi \geq F_A^*) + 1}{(\#\ permutations + 1)}$$

where the 1 is included because the observed F-ratio is considered one of the many outputs that might have occured under the null hypothesis and was the correction proposed by Dwass, M. (1957) in order to preserve the type I error in the Monte Carlo approach.

2.3.3 Assumptions and characteristics

2.3.3.1 Exchangeability The only assumption of the PERMANOVA procedure is exchangeability, which is achieved if observations are independent and have similar dispersions, as was explained in Subsection 2.2.

2.3.3.2 Heteroscedasticity The PERMANOVA procedure is sensible to departures from the assumption of homoscedasticity. When homoscedasticity is not fulfiled, then the rejection of the null hypothesis of no differences in locations might actually be due to differences in dispersions.

2.3.3.3 Correlations The pseudo F-ratio of PERMANOVA doesn't consider the correlation among variables, so in this sense is robust against heteroscedasticity in the covariances. A parametric statistic to test in one-way MANOVA such as Wilks' Lambda is based on information abouts sum of squares and cross products. Indeed, let $\boldsymbol{B}$ be the matrix which collects sums of squares and cross products between groups:

$$\boldsymbol{B} = \sum_{i=1}^{g} n_i(\bar{\boldsymbol{y}}_{i.} - \bar{\boldsymbol{y}}_{..})(\bar{\boldsymbol{y}}_{i.} - \bar{\boldsymbol{y}}_{..})'$$

where elements of the diagonal are the sum of squares while the off diagonal elements are the cross products and i refers to the group. Similarly, within-groups information is

collected in the matrix

$$\boldsymbol{W} = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_{i\cdot})(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_{i\cdot})'$$

where j refers to a replicate in a given group i. Then Wilks' lambda is

$$\Lambda = \frac{|\boldsymbol{W}|}{|\boldsymbol{E} + \boldsymbol{B}|}$$

which takes into consideration both sum of squares and cross products. The pseudo F-statistic, in considering the trace of such an outer product, doesn't take into consideration the correlation among variables.

3 Simulations

3.1 Design

3.1.1 Purpose

The PERMANOVA procedure is designed for testing differences among means, the multivariate centroids among the groups. Stability of dispersions among groups allows the test to identify the difference due to the location with more accuracy. As already mentioned in previous sections (Aderson, M.J., 2001), the PERMANOVA procedure is sensible to differences in dispersions: it becomes more difficult to decide whether a difference among groups is due to the central locations when there are also different dispersions affecting the data.

In this section several simulation analyses were performed in order to check empirically, under certain conditions, how sensible the procedure might be to different dispersions among groups of interest when the null hypothesis is true. Here is important to distinguish between variance and dispersion. The former is one of the possible sources of dispersion and refers to a parametric aspect of a multivariate distribution, for example the variance-covariance values of a multivariate normal distribution. Dispersion is an empirical aspect of data that might arise also under the data's layout conditions, for example unbalancedness in the factors' combinations or empty cells, which is a severe case of unbalancedness (Searle, 1987).

It was indeed the case of ecological data reported in Section 4, which inspired us to perform a simulation study of the effects that unbalancedness and especially empty cells can have on the exactness of the PERMANOVA test, not only different variances.

The target of the simulations was the real rate of rejections of the null hypothesis, assuming that it was true. This indicates, heuristically, how conservative (lower than the *a priori* established level of significance) or liberal (higher than the *a priori* established level of significance) the test might perform in cases of severe unbalancedness and unequal dispersions. Unbalancedness includes not only the case of unequal sample sizes of observations across each factors' levels, but also the case of empty cells, which corresponds to combinations of factors levels whose information is missing. In the real case commented in Section 4, there are no available data of nutrients for June 2012 under the presence of seabirds nesting colonies (see Table 12).

3.1.2 Montecarlo simulations

The empirical study was performed by means of Monte Carlo simulations - in this case by means of repeatedly sampling from known distributions. The procedure allows to check the above mentioned robustness by controlling for the null hypothesis: the random variates were simulated under a true null hypothesis of equal means or central locations in each scenario.

All cases corresponded to a fixed design of two factors A and B, with 2 levels in factor A and 3 levels in factor B. This decision was only conventional and due to computational limitations. It didn't correspond to experimental purposes, as it might have been also a 1 factor design with g levels. As a further development, several scenarios might be built each one corresponding to a complex experimental design, allowing also for different numbers of levels in each factor. The implementation was permformed in the R environment and as far as we know the only package designed for PERMANOVA analyses is *Vegan* (Oksanen, J. et al. 2001). Function *adonis* runs the permutational MANOVA test and is of direct use for fixed factors designs. Including mixed or pure random designs would have required further computational modifications of the code that were beyond the purposes of this work.

3.1.3 Simulation parameters

In an empirical simulation study, a *scenario* refers to a set of parameters under which a data set is generated. In the present study, each data set consisted of a multivariate layout of $N \times p$ values, where N is the size of the whole sample of observations and p the length or number of components of each observation. Each layout was multivariate since p was always greater than 1. In this section all the relevant parameters are introduced.

3.1.3.1 Variates and heteroscedasticity Each simulated data was generated either from a multivariate normal distribution without correlation structure -independent observations- or from a negative binomial. The later, to consider count data, very frequent in ecology. The consideration of zero correlation was only conventional but partially supported by considerations already exposed in Paragraph 2.3.3.3. Simulating several correlation structures might be indeed a further development.

The parameters governing the multivariate normal are a vector of means $\boldsymbol{\mu}$ and a matrix of variance-covariances which, under zero correlation reduces to a diagonal matrix of variances. Under a true null hypothesis all vector's means are considered equal. Dispersion through inequality of variances is simulated in relation to the levels of factor A. For an N total sample size of simulated variates, N_1 samples correspond to level 1 of factor A and N_2 to level 2 (see Table 1). From now on, dispersion and heteroscedasticity will be considered as exchangeable terms.

Variance affecting the N_2 samples of level 2 of factor A, Σ_2, is forced to be different from Σ_1 by means of a multiplier k which increasingly adds more dispersion in N_2 in relation to N_1 according to the simple rule:

$$\Sigma_2 = k\Sigma_1$$

Eventhough Σ_1 affects only N_1, samples corresponding to the B factor are affected by both variances. For example N_{11} and N_{21} are those samples corresponding to level 1 of factor B and N_{11} is affected by Σ_1 and N_{21} by Σ_1. This is a way of varying dispersion among levels of A by somehow *keeping constant* dispersions among the levels of factor B.

Variance	levels A	N_A	levels B	N_B
Σ_1	1	N_1	1	N_{11}
			2	N_{12}
			3	N_{13}
$\Sigma_2 = k\Sigma_1$	2	N_2	1	N_{21}
			2	N_{22}
			3	N_{23}
Totals		N		N

Table 1: Variance structure for each simulated sample size. Σ_1 is the unit variance matrix governing the distribution of each observation in N_1 while Σ_2 is the variance matrix governing the distribution of each observation in N_2. Each N_{ij} refers to a sample size belonging to the i^th level of factor A and the j^th level of factor B. Finally k is the scalar that amplifies the variance in N_2 in relation to the variance in N_1.

In order to account for count data, variates from the negative binomial distribution were also simulated. The *rnegbin* function of the *MASS* package in R was employed, according to which the central location is ruled by a μ parameter. The variance σ obeys to the formula

$$\sigma = \mu + \frac{\mu^2}{\theta}$$

were θ is a measure of the shape. By decreasing the value of θ the variance will increase, keeping constant the mean. So, even though the variance might increase with the mean, which doesn't happen in the case of the normal distribution, it is possible to control for the variance independently of μ. In order to allow for N_2 to have a different variance from N_1, θ_2 was set to $\theta_2 = v\theta_1$. By means of further decreasing the value of v, σ_2 was forced to be higher than σ_1. Since no correlation structure was considered, the multivariate observations were built by means of binding univariate negative binomial variables independently simulated. As an example in 2 dimensions, Figure 4 *b* shows a bivariate negative binomial distribution of two independent simulated random variables with greater dispersion that the one in case *a* by controlling the parameter θ.

3.1.3.2 Unbalancedness As already mentioned, the present empirical study was motivated by real data with severe unbalancedness. Zero-cells data is a special case of unbalancedness and according to referencial bibliography it demands always special attention when it comes to interpret results of an analysis of variance based on sums of squares (Searle, S.R., 1987).

Unbalancedness was induced randomly by means of a multinomial distribution where the complete set of cells of the 2×3 design (see Table 2) corresponded to the realization of a multinomial process by keeping constant the total simulated sample size N. Unbalancedness, included the empty-cell case, was induced by means of controlling the vector of probabilities governing the multivariate distribution. Of course, other methods might have been possible, as soon as unbalancedness can be induced in a way that reflects dif-

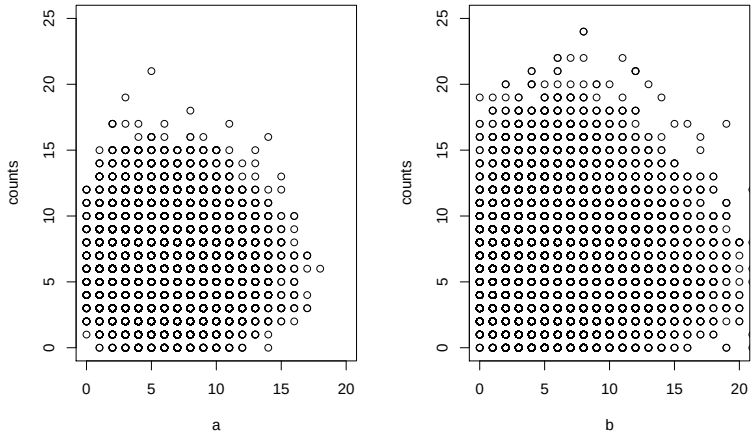

Figure 4: Two bivariate negative binomial distributions. Case b shows a greater dispersion by multiplying θ_1 by a scalar smaller than 1.

ferent degrees of heterogeneity.

		Factor B		
		1	2	3
Factor A	1	N_{11}	N_{12}	N_{13}
	2	N_{21}	N_{22}	N_{23}

Table 2: Grid of sample sizes per factor combination.

As an example, Tables 3 to 5 show three cases of unbalancedness: Table 3 is the complete balanced case with all counts of samples per factor combination equal for a total N of 120. Table 4 shows the case of unbalancedness in which there are unequal $N_{ij}s$ to estimate the means but all $N_{ij}s$ are available and $\sum_{i=1}^{2}\sum_{j=1}^{3} N_{ij} = N$. Finally Table 5 shows the case of unbalancedness in which there is missing information for the factorial combination $A_1 \times B_1$ eventhough still $\sum_{i=1}^{2}\sum_{j=1}^{3} N_{ij} = N$. The constraint that $\sum_{i=1}^{2}\sum_{j=1}^{3} N_{ij} = N$ is pure conventional with the aim to control for the total sample size while inducing heterogeneity in the subsamples N_{ij}. Of course nothing avoids to allow for simultaneously varying both the heterogeneity and the total sample size.

	1	2	3
1	20	20	20
2	20	20	20

Table 3: Complete balanced: equal sample sizes and complete cells.

	1	2	3
1	19	21	14
2	24	28	14

Table 4: Unequal sample sizes with complete cells.

	1	2	3
1	0	21	25
2	24	32	18

Table 5: Unbalanced with zero cells: no N_{11} observations available for the (11) factor combination.

3.1.3.3 Distance Since the analysis of variace is performed on the scale of the resemblance measures, the choice of it is crucial since a given measure might transform the values in a way that could indeed hide the difference of centroids among factors levels: the Bray Curtis resemblance measure doesn't conserve the location difference that might be present in the Euclidean space of the original data (Anderson, M. J. and Walsh, D.C.I., 2003).
In the present case the Euclidean distance measure was used for both the multivariate normal and negative binomial sample units. Since the Bray Curtis' distance is widely used in community data, this was also employed in the simulations only over the negative binomial samples.

3.1.3.4 Sample size and variables Total sample size was allow to vary across twwo levels, at $N = 60$ and $N = 120$. It was indeed of interest to observe whether the real rate of rejections reached the nominal when N increased. A special attention deserves p, the number of variables in each observation. It is known that i) the parametric test MANOVA is not operative in the case of $p > N$ and that ii) in ecology data it is frequent to find situations in which many variables are collected for a few number of sample units. Considering this, it was also of interest to examine situations in which $p > N$.

3.1.3.5 Permutations and replicates D, the number of replicates, controls how many multivariate data sets are generated for each setting of parameters or scenario. Each scenario is executed D=1000 times in order to have a proper number of Monte Carlo data sets on which to apply a PERMANOVA test. Each of them employs a number P of unrestricted permutations in order to estimate a P value.

3.1.4 Scenarios

Table 6 summarizes all the examined values which will be commented in the present section. As already mentioned, variates where sampled D times from the multivariate normal distribution and the negative binomial, both without any correlation structure. Each data sample consisted of N units of p variables, a layout of $N \times p$ data.

There are several factors that play a role -total sample size, number of variables, distance, dispersion and so on- so it is mandatory to structure the analysis in some way. We considered reasonable to divide the scenarios into two big groups, those in which $N < p$ (Table 7) and those in which $N > p$ (Table 8). The logic behind this relies on the frequent situation in ecology, in which there are meany measurements or characteristics available for a reduced sample size. Technology allows for many enviromental data retrieves and at the same time the fieldwork might allow only for a scarse number of sample units. Another reason relies on the PERMANOVA method itself and its capability to handle cases in which $N < p$, whereas the parametric version MANOVA does not. Nevertheless it is important to note that *a priori* no particular effect on real rejections was expected from different settings of the ratio N/p. This was so because the PERMANOVA test works on the distance between objets, either these are less or more than the number of variables.

In the first group of simulations p was set at 4 in combination of 60 and 120 sample units (Table 7). The means of the multivariate normal and the negative binomial were set at $\mu = (6, 6, 6, 6)$ as a pure convention. To examine a more realistic scenario in the case of ecology data, p was set at 150 in the second group (Table 8), greater than the number of sample units in both the scenarios where $N = 60$ and $N = 120$, by simulating variates with μ vectors of $\mathbf{6}_{150}$.

These combinations of N and p were simulated under the Euclidean and Bray-Curtis resemblance measures, the second one applied only to the negative binomial distribution, the count data. This means that each original $N \times p$ data set was, following the PERMANOVA procedure, transformed to the scale of the resemblance measure and then each distance value labeled according to the factor combination it belonged to.

Several situations of balancedness/unbalancedness were considered:

1. *B (with no empty cells).* It corresponds to the case of Table 3 with equal sample sizes $N_{ij}s$ and complete information (no empty cells). This is the easiest case when it comes to interpret the results of an analysis of variance by means of the sums of squares;

2. *B (with 1 empty cell).* It corresponds to a grid with equal sample sizes $N_{ij}s$ with the exception of one $N_{i'j'}$ empty cell and the rest, balanced. It is like the case in Table 3 but with one cell empty. All the rest of cells sum up to N;

3. *B (with 2 empty cells).* It represents the same case as the previous one but with two empty cells, with more missing information;

4. *U (with no empty cells).* It corresponds to the case of Table 4 in which the ratios among the cells counts are different to 1 but no information is missing regarding the factor combinations;

5. *U (with no empty cells, $N_{2.} > N_{1.}$).* It corresponds to a situation of heterogeneity in cells counts but it was unbalanced on purpose in a way that each cell belonging to the level 2 of factor A has more sample units than each cell of level 1. As a

consequence, the total sample size of level 2 of factor A is greater than the total sample size of level 1 of the same factor. Furthermore, there are no empty cells;

6. *U (with no empty cells, $N_{2.} < N_{1.}$).* It corresponds to a situation of heterogeneity in cells counts but it was unbalanced on purpose in a way that each cell belonging to the first level of factor A has more sample units than each cell of the second level. As a consequence, the total sample size of level 2 of factor A is smaller than the total sample size of level 1 of the same factor. Furthermore, there are no empty cells;

7. *U (with 1 empty cell).* It corresponds to a case of unequal sample sizes and one empty cell;

8. *U (with 2 empty cells).* It corresponds to a case of unequal sample sizes and 2 empty cells;

9. *U (with 2 empty cells).* It corresponds to a case of unequal sample sizes and 3 empty cells.

For each N/p ratio (the two groups as above commented), and for each distance, distribution function and unbalancedness structure, 2 sets of 4 data sets were simulated, each one of them corresponding to a variance structure for a total of 432 simulations as detailes in Tables 7 and 8. Following the structure of Table 7, the first 4 scenarios for $N = 60$ were built on the following settings:

1. $N = 60 > p = 4$;

2. *Distance*: Euclidean;

3. *Distribution*: Multivariate normal (MVN);

4. *Unbalancedness*: B(with no empty cells), so completely balanced;

5. D=1000 and P=999.

and setting the value of k to 1, 3, 10 and 20.

As another example, the 4 scenarios corresponding to the entry 23 and $N = 120$ in Table 8 were built on the following settings:

1. $N = 120 < p = 150$;

2. *Distance*: Bray Curtis;

3. *Distribution*: Negative binomial (Nbin);

4. *Unbalancedness*: U (with no empty cells and $N_{2.} > N_{1.}$);

5. D=1000 and P=999.

and setting the value of v to 1, 0.05, 0.02 and 0.01.

In the study, dispersion among groups was intended as the data corresponding to one level of factor A having one variance and the other level having the same one multiplied by a scalar that increases it. Being Σ_1 the unit variance matrix (of $p \times p$) of the multivariate normal distribution, with zero correlation, k was set up at 1 in order to allow for equal dispersion and also at 3, 10 and 20, as follows:

$$\Sigma_2 = k \times \Sigma_1 = k \times \mathbf{1}_{p \times p} \text{ for } k = 1, 3, 10, 20$$

In an analogous way, the variance of the p vectors of negative binomial observations were controlled by the shape parameter θ as already above explained. Given that $\sigma = \mu + \frac{\mu^2}{\theta}$, increasing θ decreases the variance up to a value near to the mean when it tends to infinity. The control scalar v was set up at 1, for the case of equal θ_s and variances as well at 0.05, 0.02 and 0.01 each one of them allowing for an increasing variance.

$$\sigma_2 = \mu + \frac{\mu^2}{v \times \theta_1}$$

The values of v were chosen by a descriptive way given $\mu = 6$. A given $\theta = 500$ gives a variance $\sigma = 6.072$. As can be seen in Figure 5 further decreasing v from 0.4, more or less, starts to give values of the variance for the second group that are reasonable to simulate a different dispersion.

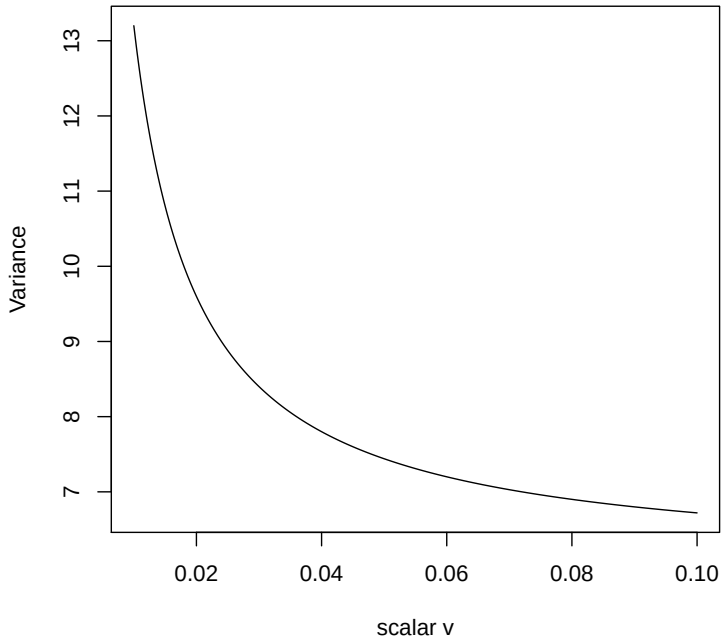

Figure 5: Standard deviation as a function of scalar v for a given $\mu = 6$ and $\theta = 500$.

Parameter	Definition	Possible values
Distribution	Distribution function from which the independent variates where sampled	Multivariate Normal (MVN) , Negative Binomial (Nbin)
N	Total sample size	60, 120
p	Number of variables in each observation	4, 150
μ_{MVN}	Mean parameter of length p	p times the same value, 6
Σ_1	Variance matrix of the multivariate normal	Diagonal unit matrix of size $p \times p$
k	Scale parameter which increases Σ_2 according to $\Sigma_2 = k\Sigma_1$	1,3,10,20
μ_{NB}	Mean parameter	p times the same value, 6
θ_1	Shape parameter of the negative binomial variates, which allows to control the variance σ according to $\sigma = \mu + \frac{\mu^2}{\theta}$	p times the same value, 500
v	Scale parameter which increases the variance of the negative binomial variates	1, 0.05, 0.02, 0.01
P	Permutations in each PERMANOVA test	999
D	Number of Monte Carlo $N \times p$ data sets generated	1000
Distance	Resemblance measure	Euclidean, Bray-Curtis
Unbalancedness	The level of balancedness in the count of sample units across all factorial combinations	B (equal sample sizes, including the case of empty cells) and U(unequal sample sizes, including the case of zero cells)

Table 6: Scenarios' parameters.

3.2 Results

3.2.1 Robustness against dispersions when $N > p$

3.2.1.1 Multivariate normal variates and euclidean distance Dispersion as due to increasing different vaiances was compared on each of the scenarios as summarized in Tables 7 and 8. Figure 6 summarizes graphically the results corresponding to the first block of scenarios, rows 1 to 9 of Table 7, for the group of simulations in which:

- $N > p$ (Table 7);
- variates were simulated as multivariate normal distributed;
- the distance used was Euclidean.

	N/p ratio	Distance	Distribution	Unbalancedness	Dispersion	Cases (N=60)	Cases (N=120)
1	$N > p = 4$	Euclidean	MVN	B (with no empty cells)	k = 1,3,10,20	4	4
2	$N > p = 4$	Euclidean	MVN	B (with 1 empty cell)	k = 1,3,10,20	4	4
3	$N > p = 4$	Euclidean	MVN	B (with 2 empty cells)	k = 1,3,10,20	4	4
4	$N > p = 4$	Euclidean	MVN	U (with no empty cells)	k = 1,3,10,20	4	4
5	$N > p = 4$	Euclidean	MVN	U (with no empty cells $N_{2\cdot} > N_{1\cdot}$)	k = 1,3,10,20	4	4
6	$N > p = 4$	Euclidean	MVN	U (with no empty cells $N_{2\cdot} < N_{1\cdot}$)	k = 1,3,10,20	4	4
7	$N > p = 4$	Euclidean	MVN	U (with 1 empty cell)	k = 1,3,10,20	4	4
8	$N > p = 4$	Euclidean	MVN	U (with 2 empty cells)	k = 1,3,10,20	4	4
9	$N > p = 4$	Euclidean	MVN	U (with 3 empty cells)	k = 1,3,10,20	4	4
10	$N > p = 4$	Euclidean	Nbin	B(with no empty cells)	v = 1,0.05,0.02,0.01	4	4
11	$N > p = 4$	Euclidean	Nbin	B (with 1 empty cell)	v = 1,0.05,0.02,0.01	4	4
12	$N > p = 4$	Euclidean	Nbin	B (with 2 empty cells)	v = 1,0.05,0.02,0.01	4	4
13	$N > p = 4$	Euclidean	Nbin	U (with no empty cells)	v = 1,0.05,0.02,0.01	4	4
14	$N > p = 4$	Euclidean	Nbin	U (with no empty cells $N_{2\cdot} > N_{1\cdot}$)	v = 1,0.05,0.02,0.01	4	4
15	$N > p = 4$	Euclidean	Nbin	U (with no empty cells $N_{2\cdot} < N_{1\cdot}$)	v = 1,0.05,0.02,0.01	4	4
16	$N > p = 4$	Euclidean	Nbin	U (with 1 empty cell)	v = 1,0.05,0.02,0.01	4	4
17	$N > p = 4$	Euclidean	Nbin	U (with 2 empty cells)	v = 1,0.05,0.02,0.01	4	4
18	$N > p = 4$	Euclidean	Nbin	U (with 3 empty cells)	v = 1,0.05,0.02,0.01	4	4
19	$N > p = 4$	Bray-Curtis	Nbin	B (with no empty cells)	v = 1,0.05,0.02,0.01	4	4
20	$N > p = 4$	Bray-Curtis	Nbin	B (with 1 empty cell)	v = 1,0.05,0.02,0.01	4	4
21	$N > p = 4$	Bray-Curtis	Nbin	B (with 2 empty cells)	v = 1,0.05,0.02,0.01	4	4
22	$N > p = 4$	Bray-Curtis	Nbin	U (with no empty cells)	v = 1,0.05,0.02,0.01	4	4
23	$N > p = 4$	Bray-Curtis	Nbin	U (with no empty cells $N_{2\cdot} > N_{1\cdot}$)	v = 1,0.05,0.02,0.01	4	4
24	$N > p = 4$	Bray-Curtis	Nbin	U (with no empty cells $N_{2\cdot} < N_{1\cdot}$)	v = 1,0.05,0.02,0.01	4	4
25	$N > p = 4$	Bray-Curtis	Nbin	U (with 1 empty cell)	v = 1,0.05,0.02,0.01	4	4
26	$N > p = 4$	Bray-Curtis	Nbin	U (with 2 empty cells)	v = 1,0.05,0.02,0.01	4	4
27	$N > p = 4$	Bray-Curtis	Nbin	U (with 3 empty cells)	v = 1,0.05,0.02,0.01	4	4

Table 7: Number of cases in each scenario structure for $N > p = 4$ (N: total sample size; p: number of variables; MVN: multivariate normal distribution; Nbin: negative binomial distribution; B: cases of equal sample sizes N_{ij}; U: cases of unequal sample sizes N_{ij}; k: parameter controlling the variance of N_2 when variates are MVN; v: parameter controlling the variance of N_2 when variates are Nbin). The number of cases in each entry is 4 since it corresponds to 4 different settings of k (1, 3, 10 or 20) or v (1, 0.05, 0.02 or 0.01) for a given distance, a given distribution and a given structure of unbalancedness in the group were $N > p = 4$.

This block corresponds to a total of 72 simulation scenarios. The vertical blue line divides the scenarios in those for which $N = 60$ and $N = 120$ (the last two columns of Tables 7 and 8). The horizontal black line corresponds to the nominal level of rejection 0.05. Each point represents a scenario simulated by $D = 1000$ Monte Carlo variates and each set of 4 points, all of them of the same color represents 4 different values of the scalar k for a given level of unbalancedness. For example the first 4 red points in figure 6 correspond to 4 sets of $D = 1000$ Monte Carlo simulations each one, all of them with level of unbalancedness *B (with no empty cells)*, which represents the complete balanced scenario (see the list above on cases of unbalancedness): for the first red point, $k = 1$, which means equality of variances, for the second one, $k = 3$, and for the third and fourth, $k = 10$ and $k = 20$ respectively. There are three sets of colors to distinguish the results and the order of plotting corresponds to the order in which are summarized in Table 7). So the first group of 4 red points corresponds to the first line in Table 7 and the case $N = 60$. The second group of blue points corresponds to the second line in Table 7 and

	N/p ratio	Distance	Distribution	Unbalancedness	Dispersion	Cases (N=60)	Cases (N=120)
1	$N < p = 150$	Euclidean	MVN	B (with no empty cells)	k = 1,3,10,20	4	4
2	$N < p = 150$	Euclidean	MVN	B (with 1 empty cell)	k = 1,3,10,20	4	4
3	$N < p = 150$	Euclidean	MVN	B (with 2 empty cells)	k = 1,3,10,20	4	4
4	$N < p = 150$	Euclidean	MVN	U (with no empty cells)	k = 1,3,10,20	4	4
5	$N < p = 150$	Euclidean	MVN	U (with no empty cells $N_{2.} > N_{1.}$)	k = 1,3,10,20	4	4
6	$N < p = 150$	Euclidean	MVN	U (with no empty cells $N_{2.} < N_{1.}$)	k = 1,3,10,20	4	4
7	$N < p = 150$	Euclidean	MVN	U (with 1 empty cell)	k = 1,3,10,20	4	4
8	$N < p = 150$	Euclidean	MVN	U (with 2 empty cells)	k = 1,3,10,20	4	4
9	$N < p = 150$	Euclidean	MVN	U (with 3 empty cells)	k = 1,3,10,20	4	4
10	$N < p = 150$	Euclidean	Nbin	B(with no empty cells)	v = 1,0.05,0.02,0.01	4	4
11	$N < p = 150$	Euclidean	Nbin	B (with 1 empty cell)	v = 1,0.05,0.02,0.01	4	4
12	$N < p = 150$	Euclidean	Nbin	B (with 2 empty cells)	v = 1,0.05,0.02,0.01	4	4
13	$N < p = 150$	Euclidean	Nbin	U (with no empty cells)	v = 1,0.05,0.02,0.01	4	4
14	$N < p = 150$	Euclidean	Nbin	U (with no empty cells $N_{2.} > N_{1.}$)	v = 1,0.05,0.02,0.01	4	4
15	$N < p = 150$	Euclidean	Nbin	U (with no empty cells $N_{2.} < N_{1.}$)	v = 1,0.05,0.02,0.01	4	4
16	$N < p = 150$	Euclidean	Nbin	U (with 1 empty cell)	v = 1,0.05,0.02,0.01	4	4
17	$N < p = 150$	Euclidean	Nbin	U (with 2 empty cells)	v = 1,0.05,0.02,0.01	4	4
18	$N < p = 150$	Euclidean	Nbin	U (with 3 empty cells)	v = 1,0.05,0.02,0.01	4	4
19	$N < p = 150$	Bray-Curtis	Nbin	B (with no empty cells)	v = 1,0.05,0.02,0.01	4	4
20	$N < p = 150$	Bray-Curtis	Nbin	B (with 1 empty cell)	v = 1,0.05,0.02,0.01	4	4
21	$N < p = 150$	Bray-Curtis	Nbin	B (with 2 empty cells)	v = 1,0.05,0.02,0.01	4	4
22	$N < p = 150$	Bray-Curtis	Nbin	U (with no empty cells)	v = 1,0.05,0.02,0.01	4	4
23	$N < p = 150$	Bray-Curtis	Nbin	U (with no empty cells $N_{2.} > N_{1.}$)	v = 1,0.05,0.02,0.01	4	4
24	$N < p = 150$	Bray-Curtis	Nbin	U (with no empty cells $N_{2.} < N_{1.}$)	v = 1,0.05,0.02,0.01	4	4
25	$N < p = 150$	Bray-Curtis	Nbin	U (with 1 empty cell)	v = 1,0.05,0.02,0.01	4	4
26	$N < p = 150$	Bray-Curtis	Nbin	U (with 2 empty cells)	v = 1,0.05,0.02,0.01	4	4
27	$N < p = 150$	Bray-Curtis	Nbin	U (with 3 empty cells)	v = 1,0.05,0.02,0.01	4	4

Table 8: Number of cases in each scenario structure for $N < p = 150$ (N: total sample size; p: number of variables; MVN: multivariate normal distribution; Nbin: negative binomial distribution; B: cases of equal sample sizes N_{ij}; U: cases of unequal sample sizes N_{ij}; k: parameter controlling the variance of N_2 when variates are MVN; v: parameter controlling the variance of N_2 when variates are Nbin). The number of cases in each entry is 4 since it corresponds to 4 different settings of k (1, 3, 10 or 20) or v (1, 0.05, 0.02 or 0.01) for a given distance, a given distribution and a given structure of unbalancedness in the group were $N < p = 150$.

the case $N = 60$. The first group of 4 red points *after the vertical blue line* corresponds to the first line in Table 7 and the case $N = 120$ (last column of Table 7). The plot at the top represents the real rejections for the A factor, while the plot at the bottom, the real rejections for the B one.

A simple view shows that the real rejections for the factor B were more stable than those for factor A, which is understandable since their levels are not affected by different variances as made on purpose on factor A. In the case of factor A, the effect of increasing variance differences among groups not always had the same effect. As already mentioned, for each group of 4 simulations of the same color, the level of unbalancedness is the same, so it is possible to see the effect of increasing heteroscedasticity by keeping constant the level of unbalancedness. Real rejections of factor A when variates are completely balanced (first 4 red points) were close to 0.05, as can bee seen from Table 25 in the Appendix: real rejections when unit variances were the same ($k = 1$) for each level of factor A and data

were completely balanced were 0.569 and increased to 0.0589, 0.0689 and 0.0649 as the value of k increased to 3, 10 and 20 respectively. Increasing the sample size to $N = 120$ made real rejections more conservative when $k = 1$ and were more close to 0.05 when k was equal to 3,10 and 20 (see first group of red points after the vertical blue line in Figure 6).

In all the rest of cases, some level of unbalancedness was introduced and real rejections increased with the level of heteroscedasticity (value of k) in the following cases:

- When counts of each factor's levels were unbalanced at random (with no structure $N_{2.} > N_{1.}$ or $N_{2.} < N_{1.}$ made on purpose) and no empty cell, which corresponds to row 4 of Table 7 or the second group of red points in Figure 6.

- When counts were unbalanced without empty cells and $N_{2.} < N_{1.}$. Note that real rejections were, on the contrary, very conservative when $N_{2.} > N_{1.}$ (row 5 of Table 7 or second group of blue points in Figure 6). As can be seen from Table 25 in Appendix, these were 0.0440, 0.0010, 0.0010 and 0.0010 when k wah equal to 1, 3, 10 and 20 respectively.

- Again when counts where unbalanced at random, and the sample size was set at $N = 120$. So it seems that having doubled the sample size didn't have any effect on improving real rejections.

- Again when counts were unbalanced without empty cells and $N_{2.} < N_{1.}$ for $N = 120$.

- Real rejections for factor B tended to increase above the nominal level, as the degree of heteroscedasticiy increased, for cases of unequal sample sizes between levels of factors (rows 6 to 9 of Table 7 for both $N = 60$ and $N = 120$)

Real rejections in factor A were found very conservative under heteroscedasticity in:

- Scenarios corresponding to rows 2, 5, 8, 9 of Table 7 when N was set at 60.
- Scenarios corresponding to rows 2, 5, 7, 8, 9 of Table 7 when N was set at 120.

Finally, Figure 6 shows that doubling the sample size from 60 to 120 had no effect on the difference between real rejections and nominal level of 0.05. Indeed the same pattern was observed and there were no bigger differences in relation to the nominal level when sample size was smaller than $N = 120$. As in Figure 6, the vertical blue line divides the experiments where $N = 60$ (left side) from those where $N = 120$ (right side). The horizontal balck line is set at 0 difference between real rejection ald nominal level 0.05.

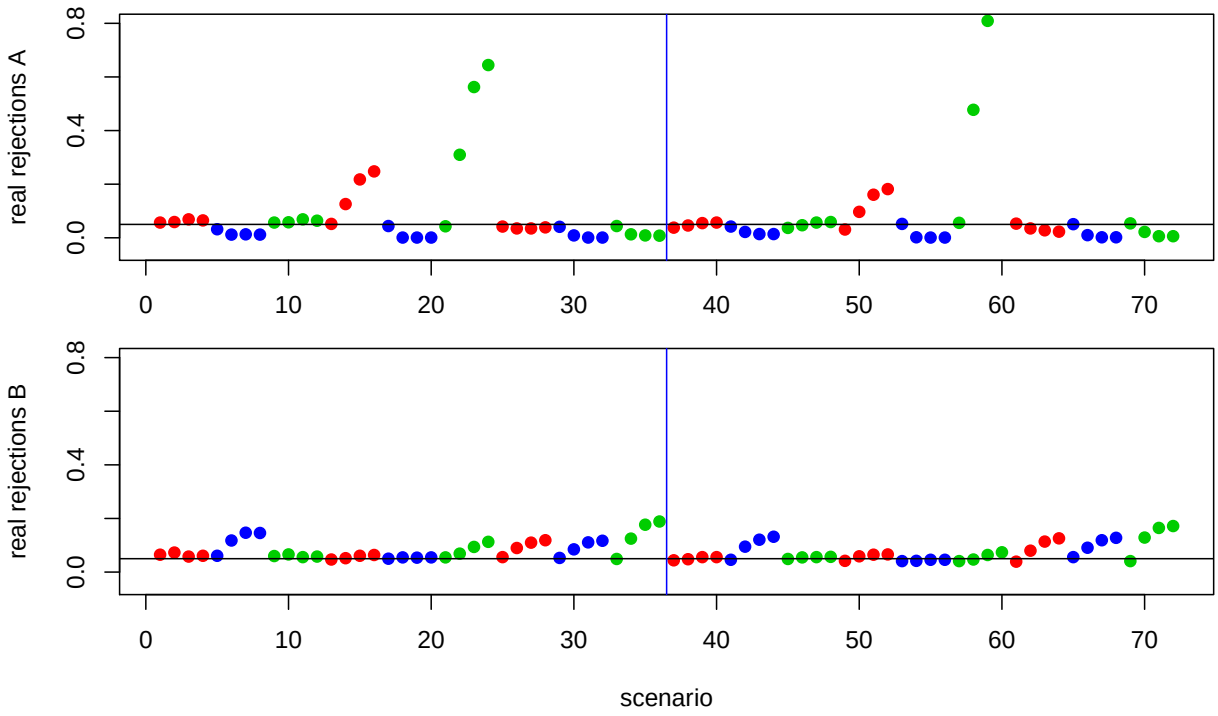

Figure 6: Real rate of rejections for cases of $N > p$ with multivariate normal variates and PERMANOVA-tested using the Euclidean resemblance measure.

3.2.1.2 Negative binomial variates and euclidean distance Figure 8 summarizes graphically the results corresponding to the second block of scenarios, rows 10 to 18 of Table 7, for the group of simulations in which:

- $N > p$ (Table 7)
- variates are simulated as negative binomial distributed
- the distance used is Euclidean.

The pattern observed was more stable around the nominal value of 0.05 than to the previous case were variates were simulated as multivariate normal.

Real rejections of factor A were found to increase with the level of heteroscedasticity (in this case as v taking the values of 1 for equal variances and the of 0.05, 0.02 and 0.01 for further inequalities of variances) when there were unequal sample sizes among levels of factors and $N_{2.} < N_{1.}$ but without empty cells, for both $N = 60$ and $N = 120$ (row 15 of Table 7).

Again, in many scenarios the behaviour of the test was found very conservative, even in cases of unbalancedness due to unequal sample sizes of counts with empty cells (cases

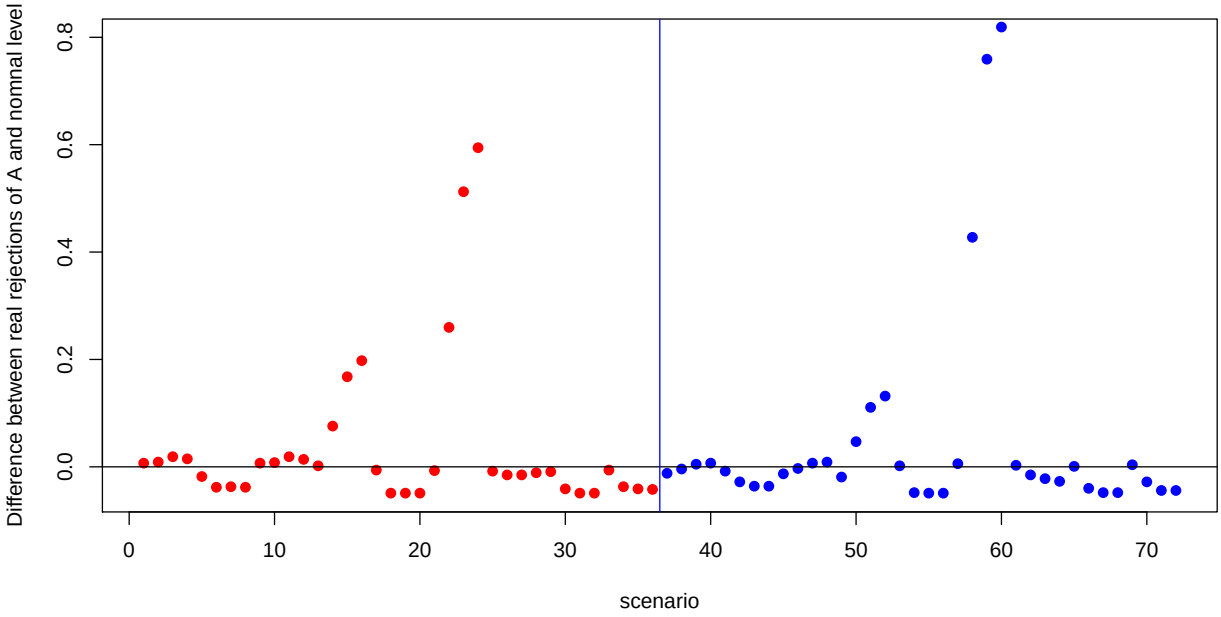

Figure 7: Difference between real rate of rejections and nominal value 0.05 for cases of $N > p$ with multivariate normal variates and PERMANOVA-tested using the Euclidean resemblance measure.

14 and 16 to 18 of Table 7 for both $N = 60$ and $N = 120$).

Real rejections in factor B were found more stable than in factor A and than in factor B of the previous scenarios. Nevertheless a slightly increased of real rejections were found as heteroscedasticity increased, especially in cases of unbalancedness due to unequal sample sizes of counts with empty cells and when N was set to 60 (as can bee seen in Table 26 in the Appendix).

Doubling the sample size barely made some difference in real rejections, with the exception of case 15 in Table 7: for $N = 120$, paradoxically, differences with the nominal level were found higher.

3.2.1.3 Negative binomial variates and Bray Curtis measure As can be appreciated from Figure 11 real rejections for both factors A and B were closed to the nominal level in almost all cases with a trend to be slightly conservative. This pattern was observed in both sample size settings, $N = 60$ and $N = 120$ (see Figure 11)

3.2.1.4 Conclusions The previous analyses concerned evaluating real rejections for increasing levels of heteroscedasticity for given levels of unbalancedness.

When $N > p$, severe differences of real rejections from the nominal value of 0.05 in factor A were observed more frequent in the case of simulated multivariate normal data

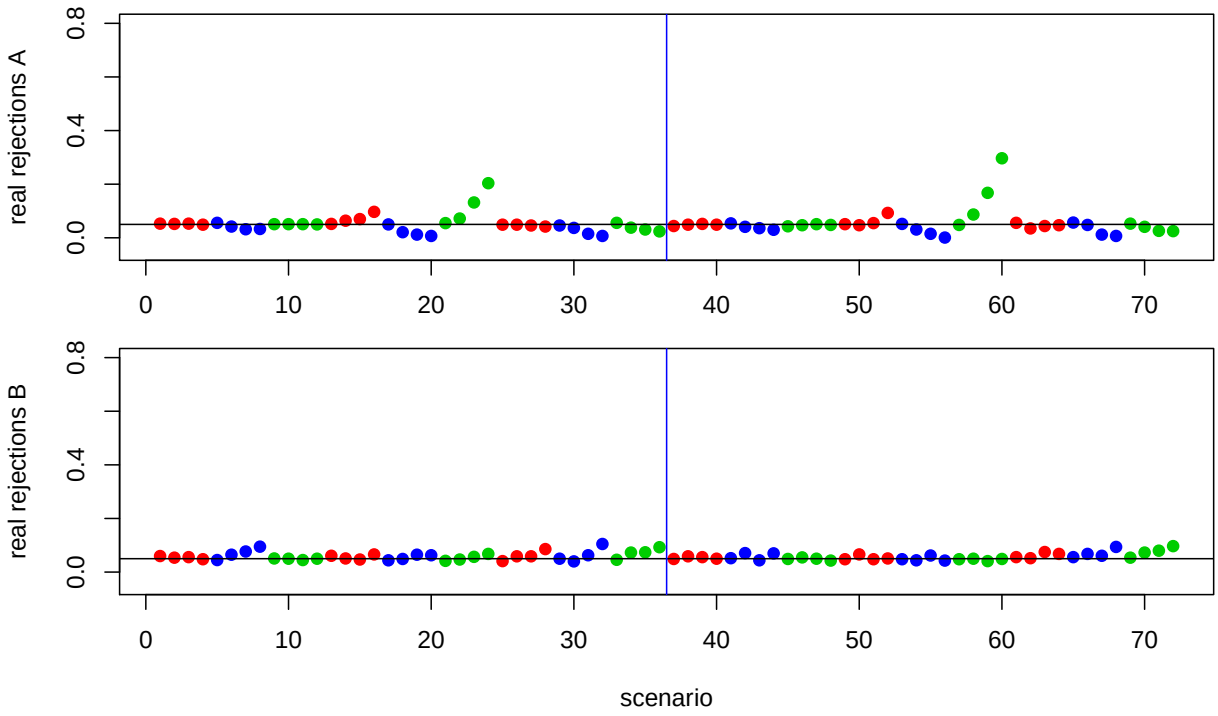

Figure 8: Real rate of rejections for cases of $N > p$ with negative binomial variates and PERMANOVA-tested using the Euclidean resemblance measure.

and when the dissimilarity measure used to build the distance matrix was the Euclidean.

Differences increased with heteroscedasticity especially when unbalancedness was due to unequal counts between factors' levels and $N_{2\cdot} > N_{1\cdot}$ (case 5 of Table 7) and unequal counts between factors' levels with 1 empty cell (case 7). This last case was also observed when data was simulated as negative binomial and the dissimilarity measure used was the Euclidean (case 14 of Table 7).

In other scenarios of unbalancedness (cases 7 to 9 and 16 to 18, where unbalancedness is due to unequal sub-sample sizes and 1 to 3 empty cells), real rejections in factor A were found conservative and showing a slightly deacreasing trend as the level of heteroscedasticity increased.

Similar trends were found for both settings of sample sizes, $N = 60$ and $N = 120$.

Real rejections in factor B showed more liberal rates as compared to factor A as well as clearer increasing trends as heteroscedasticity increased in both cases where the dissimilarity used was the Euclidean (see Figures 6 and 8) especially when unbalancedness was due to unequal sub-sample sizes and 1 to 3 empty cells (cases 7 to 9 and 16 to 18).

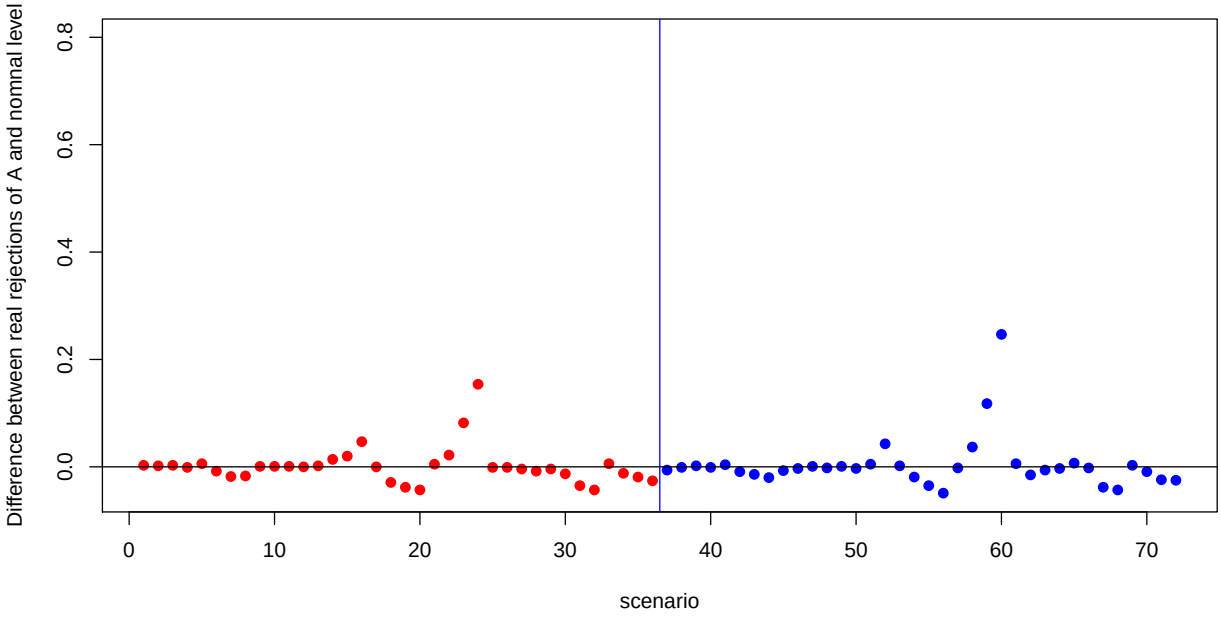

Figure 9: Difference between real rate of rejections and nominal value 0.05 for cases of $N > p$ with negaive binomial variates and PERMANOVA-tested using the Euclidean resemblance measure.

In the case of complete balance, cases 1, 10 and 19, real rejections were closed to 0.05 for both settings of sample sizes and both factors.

3.2.2 Robustness against unbalancedness when $N > p$

The empirical analysis of how unbalanced data affects real rejections is different from the analysis focused on heteroscedasticity. In the later the aim was to observe some trend as dispersion increased, either through increasing values of k or v. Instead, unbalancedness is more difficult to fit in some ordering scale. One would expect that unequal sample sizes with 2 missing cells (case 8) to be a worse scenario than unequal sample sizes without empty cells (case 4), but this would be only discretional. The reason of considering several unbalanced data was only to add some variability in the evaluated scenarios, but all cases fall into two big cathegories:

1. Complete balanced data: cases 1, 10 and 19 of Table 7, which corresponds to equal sample sizes of factors' levels and no empty cells.

2. Unbalanced data, which involves all cases different from the previous. It is worth to mention that in the interpretation of the analysis of variance results, Searle (1987) considers two subgroups of unbalanced data: those with and without empty cells.

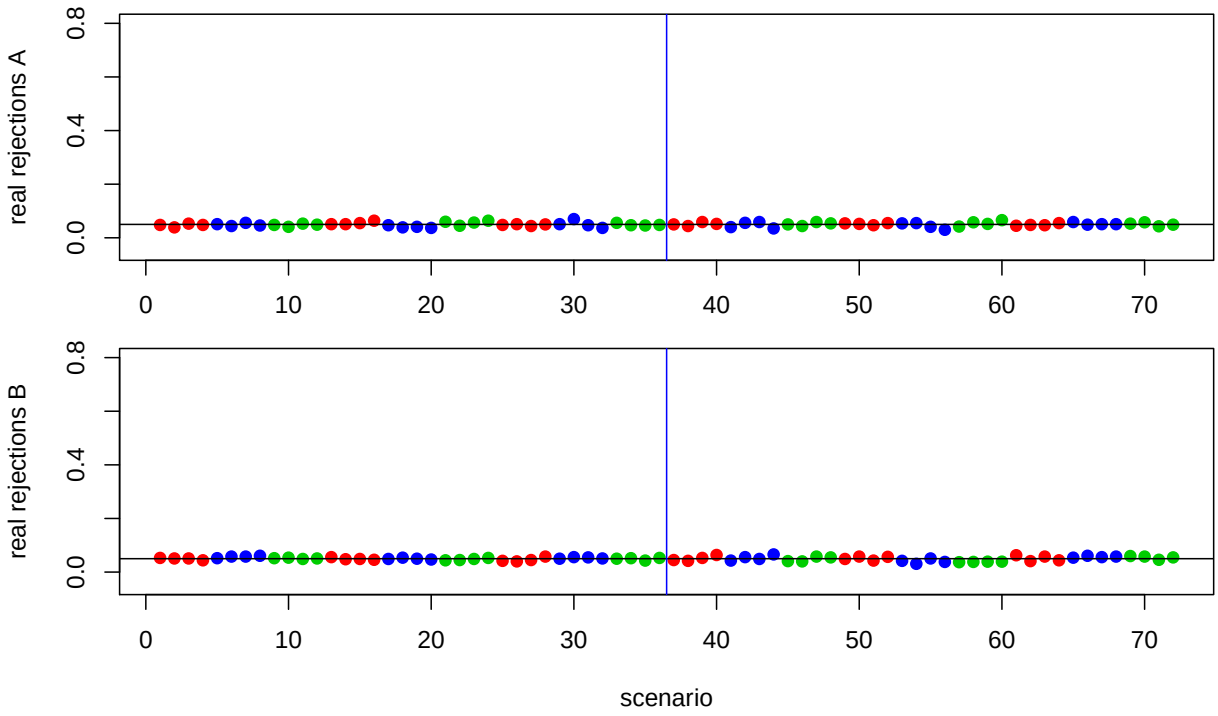

Figure 10: Real rate of rejections for cases of $N > p$ with negative binomial variates and PERMANOVA-tested using the Bray Curtis resemblance measure.

> But even if this is important in the intepretation of the results when reported as sums of squares, it is not clear whether one is a worse scenario in the empirical study.

Droping the aim of finding an incresing or decreasing trend, Tables 25, 26 and 27 in Appendix classify the empirical performance of the test into three categories based on estimated confidence intervals for the simulated real rejections:

1. If the upper limit < 0.05, then the result was considered *conservative*.
2. If the lower limit > 0.05, then the result was considered *liberal*.
3. Otherwise, was considered *closed* to the nominal or *a priori* level.

3.2.2.1 Conclusions In order to obtain a simplified vision of the test's performances, in addition to the detailed conclusions reported in the Appendix tables, some boxplots were built to reflect the empirical distribution of real rejections by splitting them into simulations corresponding to complete balanced data and to any other case of unbalancedness.

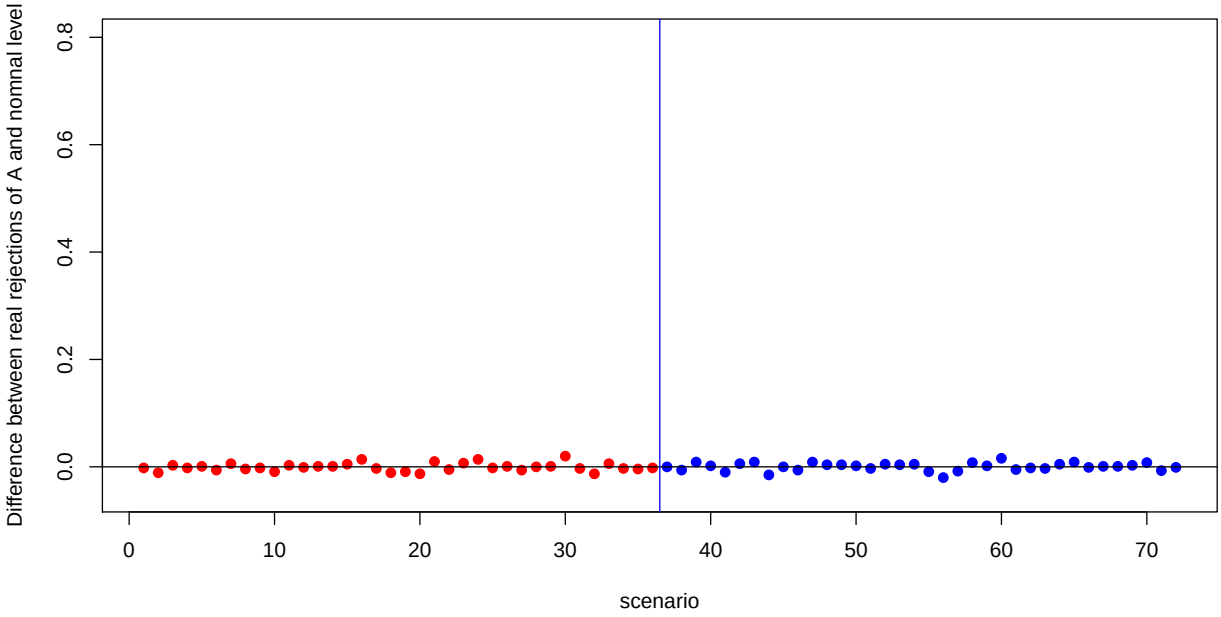

Figure 11: Difference between real rate of rejections and nominal value 0.05 for cases of $N > p$ with negaive binomial variates and PERMANOVA-tested using the Bray Curtis resemblance measure.

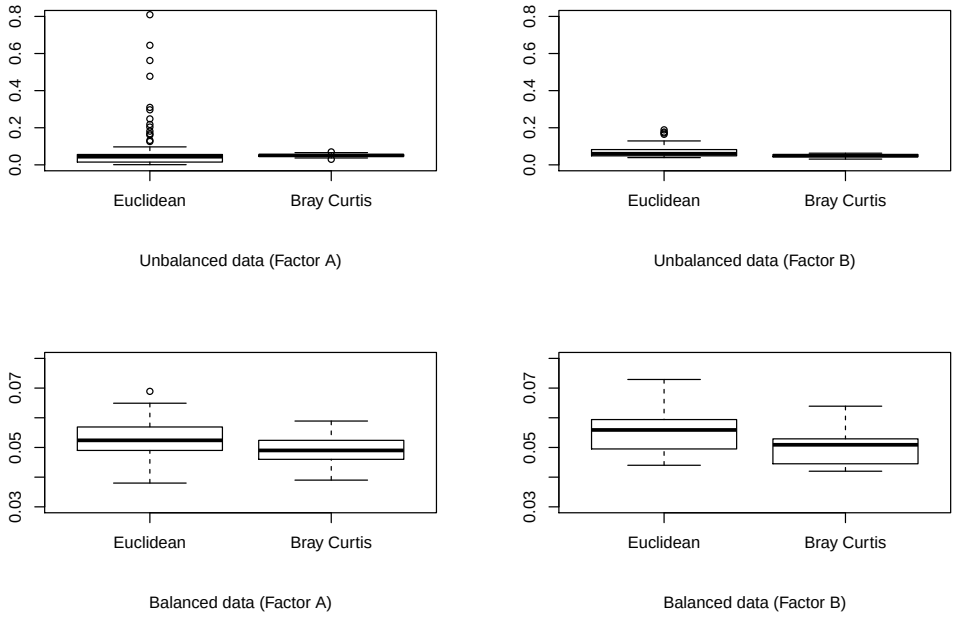

Figure 12: Observed distribution of real rejections splited by resemblance measure used to build the distance matrix: Euclidean and Bryan Curtis for count data.

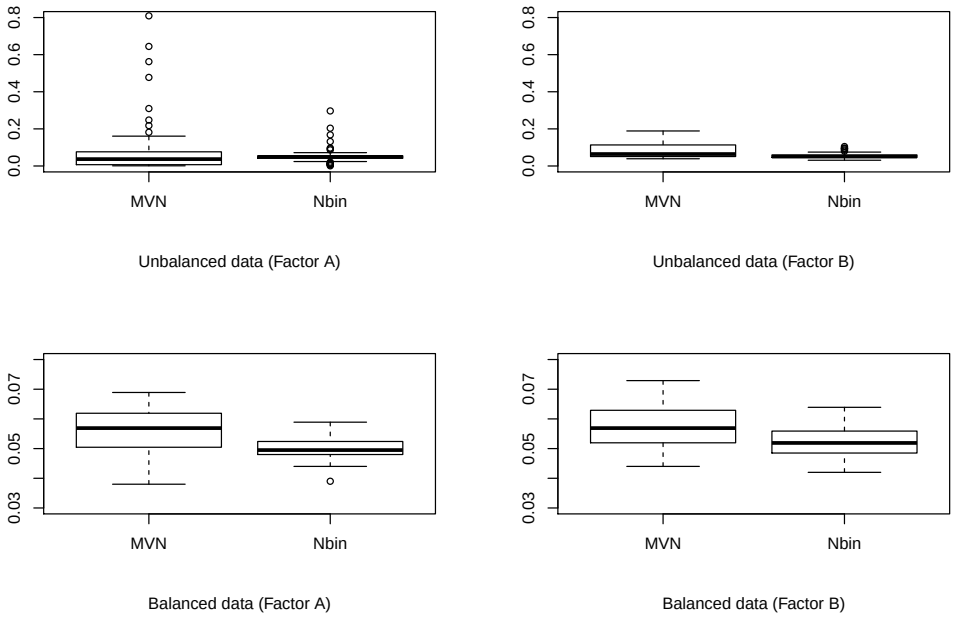

Figure 13: Observed distribution of real rejections splited by distribution function used to generate th variates: *MVN* stands for *Multivariate normal* and *Nbin* for *Negative binomial*.

- In all cases were data was balanced simulated (cases 1, 10 and 19 of Table 7), real rejections were found closed to 0.05. This case corresponds to the combination of *Balancedness=B* and *Levels=0* in Tables 25, 26 and 27 of the Appendix. Only when $k = 10$ and $N = 60$, so variance Σ_2 was 10 times unit variance Σ_1 in multivariate normal variates, the behaviour of the test was found to be liberal for complete balanced data in the case of factor A. The same behaviour was found when $k = 3$ in the analysis of factor B. When $N = 120$ the test performed closed to the nominal value in all cases of complete balanced data.

- As can be seen from Figure 12 when data were unbalanced simulated, there were some scenarios in wich the performance of the test was too liberal for factor A. The specific situations can be traced in Table 25 of the Appendix. As an example, see the case of multivariate normal variates with equal sample sizes but with 2 missing cells (*Levels=2*), $k = 10$, $N = 60$ and Euclidean resemblance measure. The performance was found liberal with a real rejection rate of 0.0689 (Table 25). These poor performances were found more frequent when the resemblance measure was Euclidean.

- The test's performance under unbalanced data was found closer to 0.05 when the Bryan Curtis measure was employed. It is worth mentioning that, nevertheless, that this resemblance index was used only in accordance with the negative binomial distribution.

- The previous conclusion in valid also when comparing the distribution models to

generate the variates: the test performed better when variates were negative binomially distributed.

- In general, for unbalanced data, real rejections were closer to 0.05 in relation to the factor B.

3.2.3 Robustness when $N < p$

The PERMANOVA test can handle data layouts, frequents in ecology and other fields, in which the number of variables is higher than the number of observations. It is still able to compute an analysis of variance since it is based on the distance matrix. This considered, the ratio N/p was not considered as an experimental factor itself, such as heteroscedasticitz or unbalancedness, but to reflect more real data contexts. Table 8 summarizes all scenarios and has indeed the same structure as Table 7 with the only difference that $p = 150$ instead of 4, which means that, for each of the distributions considered, were generated vector-variates of 150 components, more than the 60 or 120 observations available in each run.

3.2.3.1 Conclusions Figures 14 to 16 show that the effect of increasing heteroscedasticity for a given level of unbalancedness had the same trends found in the case where $N > p$, so remain valid the conclusions of the previous section.

In Figure 17 each point represents the difference between the real rejection rate for a given scenario when $p = 4$ and $p = 150$, so it is a visual tool to determine whether there are severe differences or patterns when the case $N < p = 150$ is considered instead of $N < p = 4$ and all the rest of parameters remain unchanged. The horizontal black line represents the 0 difference. As can be appreciated, increasing the number of variables up to a value higher than the number of observations, didn't add any trend to those already observed when N was set at $N > p$, but some higher dispersion can be seen in the differences related to the factor B (bottom plot of Figure 17) which means more cases in which setting $N < p$ did indeed add some noise.

In evaluating how unbalancedness affected the real rate of rejections when $N < p$, the same methods were employed, i.e. boxplots of the rejections splitted into balanced and unbalanced data as well as into distribution model and resemblance measure. These are not reported here since the observed trends were the same as those observed in Figures 12 and 13 when $N < p$.

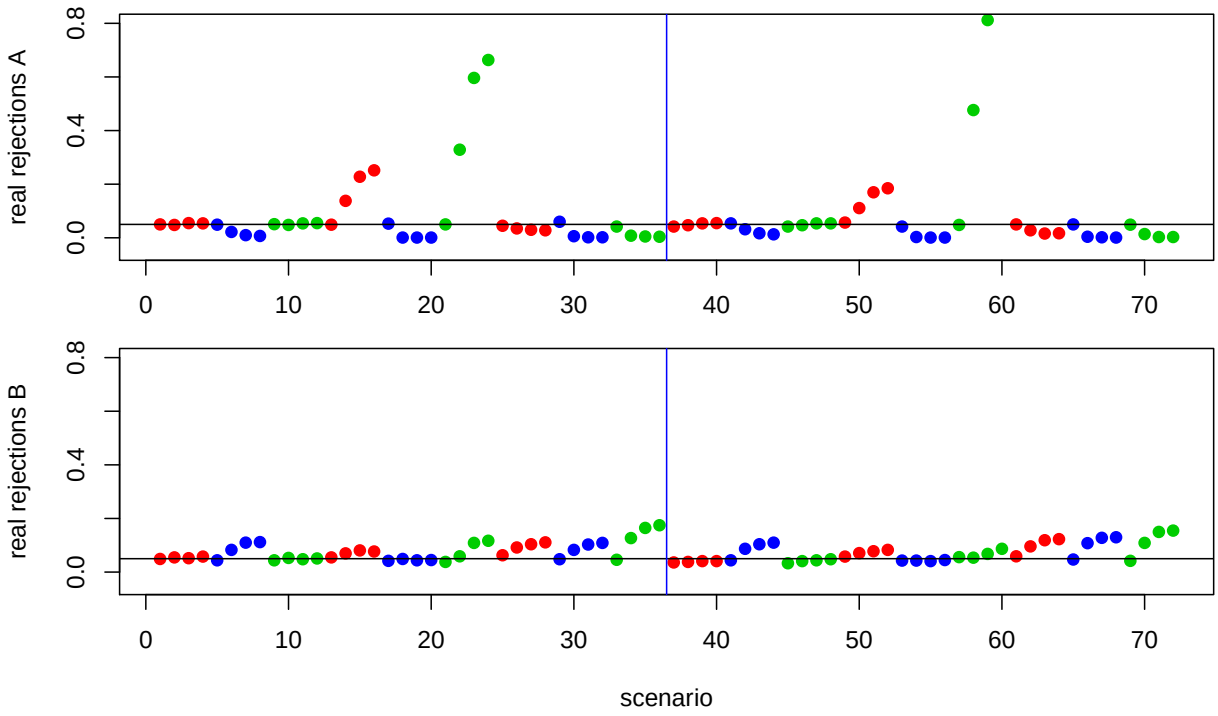

Figure 14: Real rate of rejections for cases of $N < p$ with multivariate normal variates and PERMANOVA-tested using the Euclidean resemblance measure.

Part II
Application of PERMANOVA to real data

4 Seabird nesting colonies' effects in Puerto Madryn, Patagonia

4.1 Purpose and data

4.1.1 Seabirds guano and nutrient content

Seabirds guano concentrates high levels of nutrients (nitrate, nitrite, phosphate, and ammonium) in coastal areas around their nests, providing important components such as nitrogen and phosphorous. However, high nutrient load may affect marine communities if runoff into sorrounding waters happens, as it is the case in insular and peninsular areas (Kolb, G.S. *et al.*, 2010). Researchers of the Cenpat Institut in Puerto Madryn, Argentina, have been collecting samples of water in several locations to test for the effects of seabird nesting colonies in nutrient and chlorophyll content. Data collection took place between

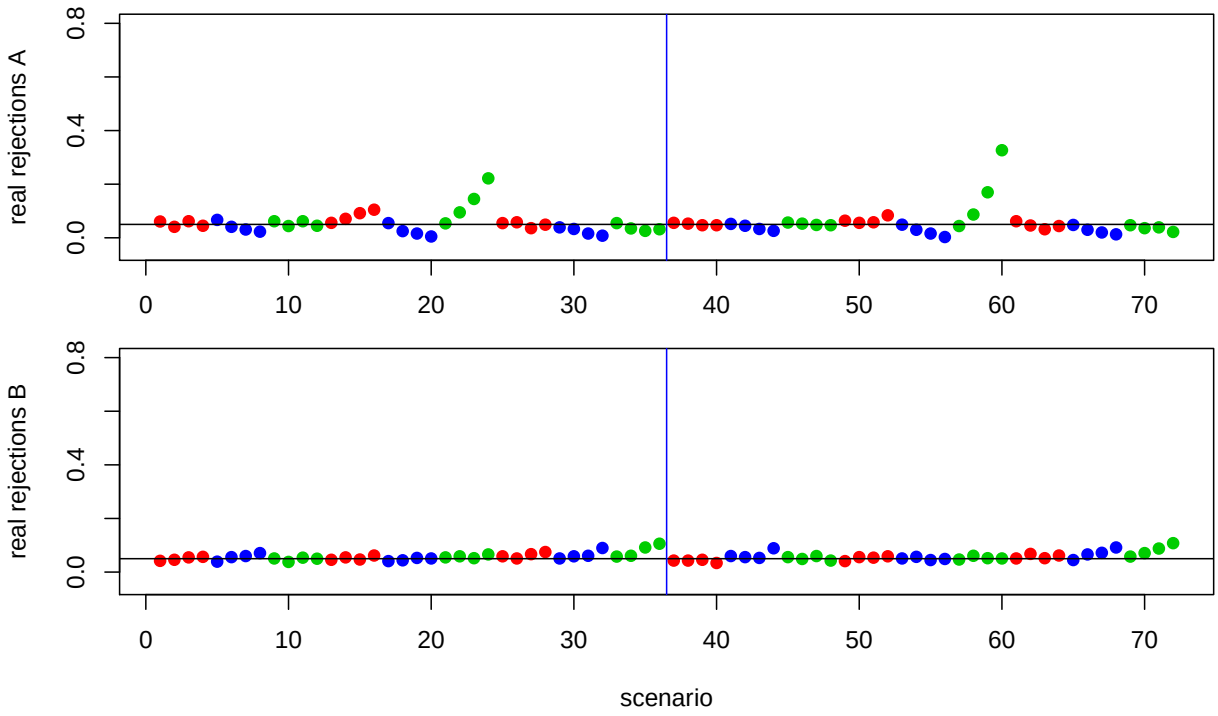

Figure 15: Real rate of rejections for cases of $N < p$ with negative binomial variates and PERMANOVA-tested using the Euclidean resemblance measure.

November 2011 and December 2012 in a coastal area near Puerto Madryn in Patagonia, Argentina. It is known by researchers that nutrients and chlorophyll follow a seasonal pattern along the year. This pattern was not of interest by the researchers, who wanted to have information on the seabird nesting colonies' presence on nutrient above and over the fluctuations due to enviromental factors, collected in a time factor. As can be seen in Figure 18 there is a high variability of chlorophyll values among periods. Not only there seem to be changes in the median but also in the dispersion -sample heteroscedasticity is a big concern for tests aimed to detect only differences between locations-. In statistical terms, the factor *period* refers to a noise that must be differentiated from the effect of nesting colonies.

4.1.2 Data

Data collection was not designed as a proper experiment due to severe limitations in reaching seabird nesting points. Periods and locations of observations were chosen depending on the presence or abscense of nesting colonies in order to build some kind of case control study. For example, June 2012, almost winter in austral hemisphere, was measured as a period without colonies, so for this month there is no crossing.

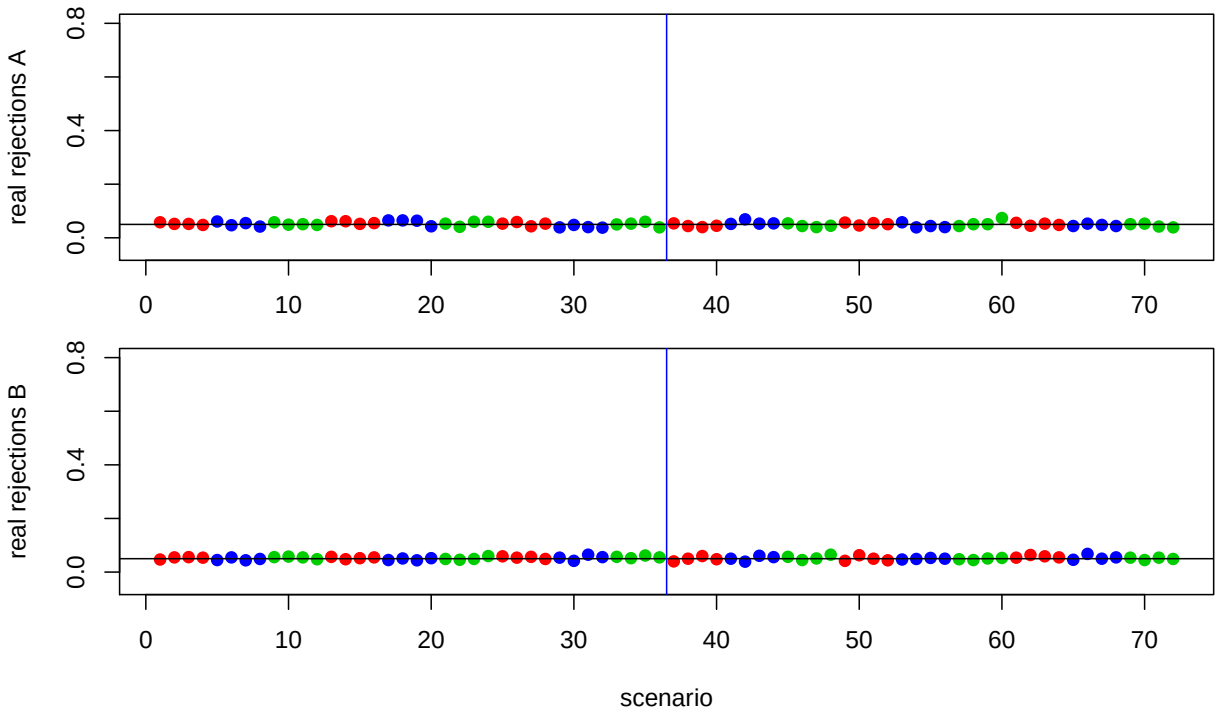

Figure 16: Real rate of rejections for cases of $N < p$ with negative binomial variates and PERMANOVA-tested using the Euclidean resemblance measure.

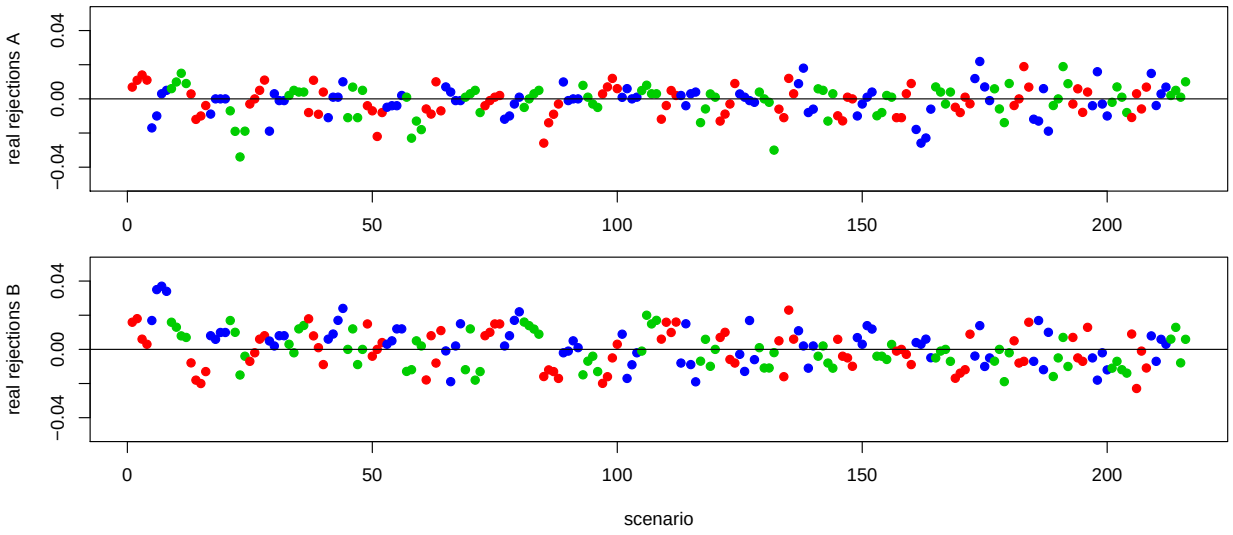

Figure 17: Difference in real rejections rate between scenarios in which $p = 4$ and $p = 150$.

4.1.2.1 Chlorophyll and nutrients Researchers took a sample of 126 experimental units, each consisting of 1 lt of sea surface water. Chlorophyll-a, from now on just chloro-

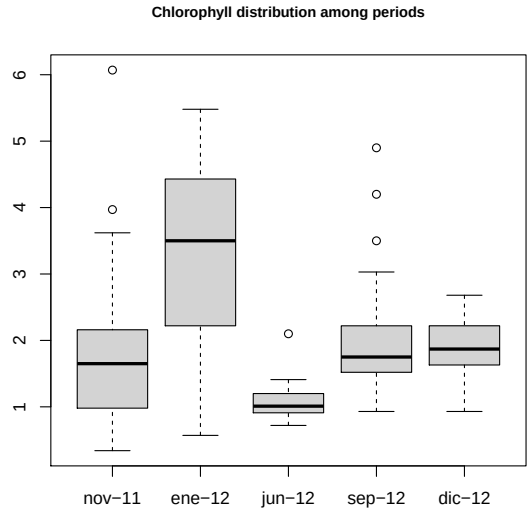

Figure 18: Distribution of chlorophyll among periods.

phyll, was measured by filtering the 1 lt samples through 47 mm Whatman GF/F filter. The filters were stored frozen at -20 C until analysis. Chlorophilly was extracted with 90% acetone and measured fluorometrically according to Strickland and Parsons (1972). Nutrients were measured on 100 ml samples taken from the filtered water, using colorimetric techniques. Table 9 summarizes each descriptor belonging to the experimental units and the unit of measurement. As can be seen, all of them have similar ranges apart from silicate which has a maximum value of 29.270. This is the only high value in the sample and after consulting with the researcher, was taken out, considered very likely a measurement error. Finally it was decided to remove all the measurements from the same experimental unit.

DESCRIPTORS FOR EACH OF THE 126 SAMPLES				
Variable	Unit of measurement	Range	Mean	Sd
Chlorophyll-a	mg/M^3	0.340 - 6.070	2.146	1.262
Nitrogen (NO2NO3)	μM	0.070 - 8.750	2.910	2.488
Phosphate	μM	0.280 - 1.320	0.853	0.262
Silicate	μM	0.830 - 29.270	2.967	2.643
Ammonium (NH4)	μM	0.030 - 3.990	0.816	0.824

Table 9: Summary of variables measured for 126 samples of water between November 2011 and December 2012. μM are micromolars.

After removing the outlier, a 125×5 multiresponse data matrix was obtained, which was analysed using the approach of the analysis of variance.

Table 10 summarizes all the variables recorded during the surveys. Not all of them were used for the statistical analysis, they just represent the raw data.

LABELS AND COVARIABLES RECORDED	
Label	Levels
Date	several dates, then transformed in *period*
Period	Nov-11 Jan-12 Jun-12 Sep-12 Dec-12
Waypoint	several levels of the kind T_{ij}
Distance to coast	a quantitative variable in meters
Seabird colonies	levels 0,1,2, then binarized into 0 (abscense), 1 (presence)
GPS point	latitude and longitude measures
Location	a collapse of *waypoint* information from T_{ij} to T_i

Table 10: Labels and covariables recorded or attributed.

4.1.2.2 Seabird colonies In order to record the factor *seabird colonies*, researchers used published information to know when and where colonies of seabirds were present. They chose the islands with the highest density and trated them as a categorical variable based on published information about the number of nests per m^2. Density can reach 2 nests per m^2. The size of the colonies depends partially on the size of the island. So according to this classification, density was recorded with possile levels 0,1 or 2 nests per m^2. Each species has a different cycle, but the main trend is that during spring (September) adults arrive at the colony and towards the end of spring they lay the eggs. After 1 month of incubation new birds are born and towards the end of March the entire colony leaves. This is the pattern for 1 species of cormorant and for penguins. There is another species of cormorants whose adults stay during the whole year. So in summer is when the abundance is the highest. From the original classification, upon consideration of the researcher, seabird nesting colonies was binarized into 0,1 for abscense and presence of seabirds colonies.

4.1.2.3 Waypoint and location *Waypoint* is a categorical variable of the kind T_{ij} which corresponds to the point of observation. The subscript i refers to the transect number (Figure 19). The set of waypoints T_{1j} and T_{6j}, belong to transects T_1 and T_6, which correspond to sites without presence of seabirds, whatever the period of the year. These sites were sampled in order to measure the nutrients under the abscense of nesting colonies.

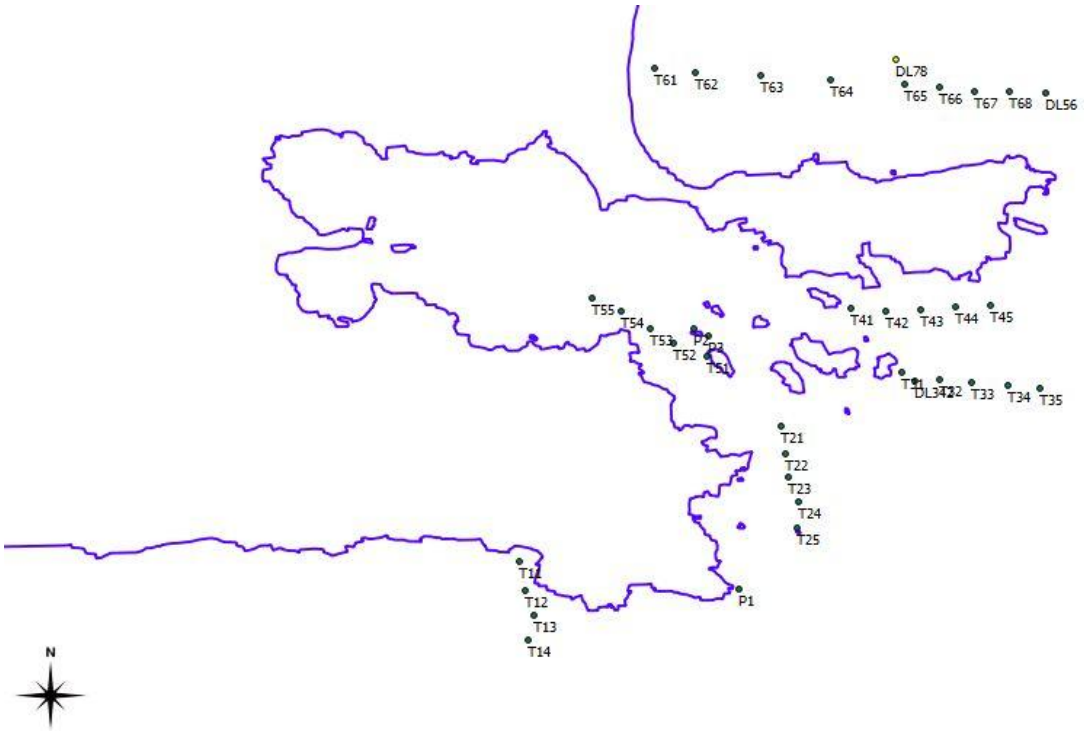

Figure 19: Sampling area.

4.1.2.4 Period As has been said, the fluctuation of nutrients along the year is considered a noise partially distorting the study of the factor of interest. *Period* doesn't cover the whole year since months were "sampled" considering the presence or abscence of seabirds and whenever was possible to reach the sampling areas.
All these considerations regarding the locations, periods and presence of nesting colonies under severe limitations of access to points of observation, had notable consequences on the design aspect of the observation, particularly on their balancedness and crossing.

4.1.2.5 Available information As mentioned, severe limitations led to unbalancedness and lack of crossing. Since the purpose was to extract useful variability information regarding the effect of nesting colonies on nutrients above and over the variability induced by enviromental variables, collected in the factor *period*, the wisdom would have been to fully sample a crossed design considering the presence or absence of colonies, the period and the location (this last two, to isolate the effect of time and space). Unfortunately this was not the case, since a 3-way design counted on few replicates, in many cases zero measures, which poses a severe limitation on having estimable parameters and functions (Searle, S.R.*et al.*, 1992 and Searle, 1987). Table 11 summarizes the available information for a 3-way crossed design. As can be seen there is not enough information for such a design. As pointed out by Searle (1987) even in case of some empty cells, the so called *geometrical-connectedness* among cells might lead to estimable parameters as functions of others, but this is not the case since it seems clear that some paths remain isolated. Geometrical connectedness for a 2-way cross design states that there is connection if "the filled cells of a grid can be joined by a continuous line, consisting solely of horizontal and vertical segments, that has changes of direction only in filled cells." (Searle, 1987, Ch. 5.3)

Since *period* still remains a factor of interest, actually a variance component, the available information for a 2-way crossed design was evaluated. Table 12 reports the number of replicates to use in a 2-way crossed design. As can be seen, there is only one empty cell, unfortunatelly belonging to *period*, June 2012, which was most sampled under abscence of seabird nesting. But *geometrical connectedness* provides a way to still have reliable estimators and sums of squares. In addition to one empty cell, there is unbalancedness. These two deficits -I consider unbalancedness a deficit in this case since is not a situation of planned unbalancedness, such as latin aquares or balanced incomplete blocks- will pose some care in the chose of which sums of squares' partition to consider in order to build a proper F-statistic.

The situation for a 2-way crossed design involving the binary variable and the locations is summarized in Table 13. It still has unbalancedness and empty cells, but the point is that, from the analysis point of view, it has more sense to model the factor of interest, nesting colonies, with a variance component such as period instead of locations, since the purpose is to infere above the variability represented by the time cycle. Space is a source of variability too, being sensible, for example, to wave exposure, which affects nutrient content even in nesting locations, but a 3-way crossed design wouldn't be reliable due to severe lack of information.

		Period	Nov-11	Jan-12	Jun-12	Sep-12	Dec-12	Sum
Seabird colonies	*Location*							
0	T1		4	4	0	4	0	12
	T2		0	0	5	0	0	5
	T3		0	0	5	0	0	5
	T4		0	0	5	0	0	5
	T5		0	0	5	0	0	5
	T6		0	8	4	5	8	25
		Sum	4	12	24	9	8	57
1	T1		0	0	0	0	0	0
	T2		5	5	0	5	0	15
	T3		5	5	0	5	4	19
	T4		5	5	0	5	4	19
	T5		5	5	0	5	0	15
	T6		0	0	0	0	0	0
		Sum	20	20	0	20	8	68

Table 11: Number of replicates available for a 3-way crossed design inolving the binary variable absence-presence of nesting colonies, month of observation and location.

	Nov-11	Jan-12	Jun-12	Sep-12	Dec-12
0	4	12	24	9	8
1	20	20	0	20	8

Table 12: Number of replicates available for a 2-way crossed design involving the binary variable absence-presence of nesting colonies, month of observation and location.

	T1	T2	T3	T4	T5	T6
0	12	5	5	5	5	25
1	0	15	19	19	15	0

Table 13: Number of replicates available for a 2-way crossed design involving the binary variable absence-presence of nesting colonies and location.

4.2 Exploratory analysis

Multivariate data is made of a 125×5 matrix of 125 samples and 5 descriptors or variables. Since all of them were taken from the same experimental unit, the approach was to consider them as a multiresponse. Nevertheless exploratory analysis was also considered for the separate distributions in order to check for individual patterns. No particular interest was posed on the kind of distribution since PERMANOVA is based on permutations and is, indeed, distribution free -free from any given mathematical known model.

4.2.1 Outliers

Is has already noted that the variable Silicate has an outlier which was its maximum value and was excluded from the data base.

4.2.2 Individual distributions

Figure 20 shows boxplots for each of the single variables. Even though they are represented in different, but similar, scales, they are still comparables in terms of dispersion, ourliers and simmetry. Considering all the values, within and outside the interquartile range, only the Phosphate seems to be not skewed, with an extreme case as NH4 which is very skewed to the left. Chlorophyll, Silicate and NH4 have also many observations falling outside the interquartile range, which is an indicator of further dispersion. Dispersion was expected since measurements were taken along several periods and in different locations. It has been showed in Figure 18 the high variability of chlorophyll all over the sampled months.

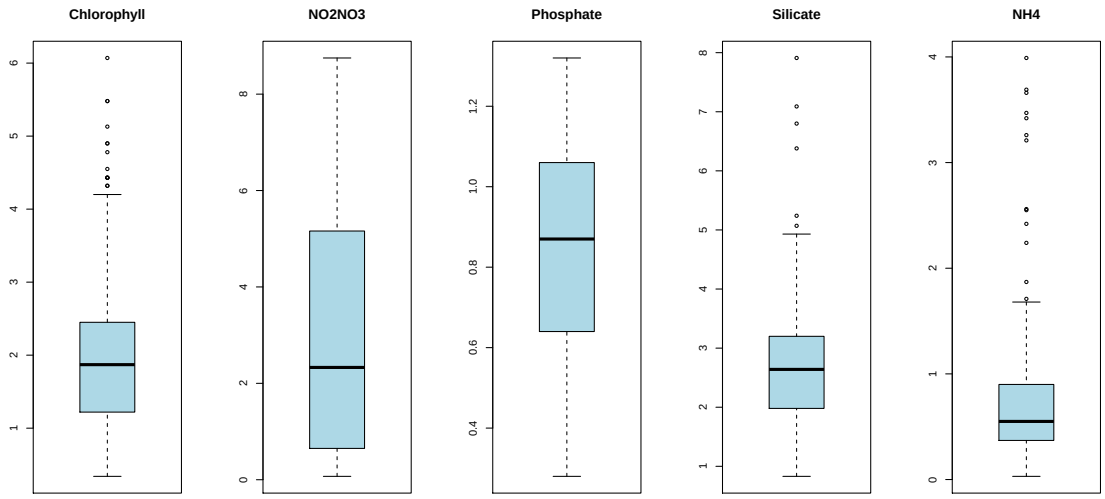

Figure 20: Sample distributions of each variable. Points repesent any univariate observation x such that $x < x_{0.25} - 1.5(x_{0.75} - x_{0.25})$ or $x > x_{0.75} + 1.5(x_{0.75} - x_{0.25})$ where each suffix indicates the order statistic.

4.2.3 Correlation among variables

The multivariate approach is usually justified when there is some correlation among components -not that much, otherwise one variable might be considered redundant-. Some correlation is expected or asummed when measurements belog to the same experimental units, as it is the case of species belonging to a community or the present case of variables measured from the same liter of water. Figure 21 shows the bivariate distributions of variables. The two colors in each plot represent abscence (black) and presence (red) of seabird colonies. Some correlation structure is clear between NO2NO3 and Phosphate. There other structures in the bivariate data eventhogh not very clearly correlated like between Silicate with Phosphate and NO2NO3. So NO2NO3, Phosphate and Silicate seem

to be more correlated among them, while NH4 and Chlorophyll seem to be more independent in relation to the rest of the variables. What seems not so clear -it was indeed tested- is the effect of the seabird guano on the bivariate distributions: the allocation seems to be random. But this was studied not on the bivariate distributions but on the multivariate distribution.
What is clear from the correlation matrix (see Table 14) is that chlorophyll is negatively correlated with all nutrients.

	chlorophyll	NO2NO3	PHOSPHATE	SILICATE	NH4
chlorophyll	1.00	-0.24	-0.28	-0.26	-0.06
NO2NO3	-0.24	1.00	0.87	0.43	0.30
PHOSPHATE	-0.28	0.87	1.00	0.42	0.29
SILICATE	-0.26	0.43	0.42	1.00	0.12
NH4	-0.06	0.30	0.29	0.12	1.00

Table 14: Correlation matrix of chlorophyll and nutrients.

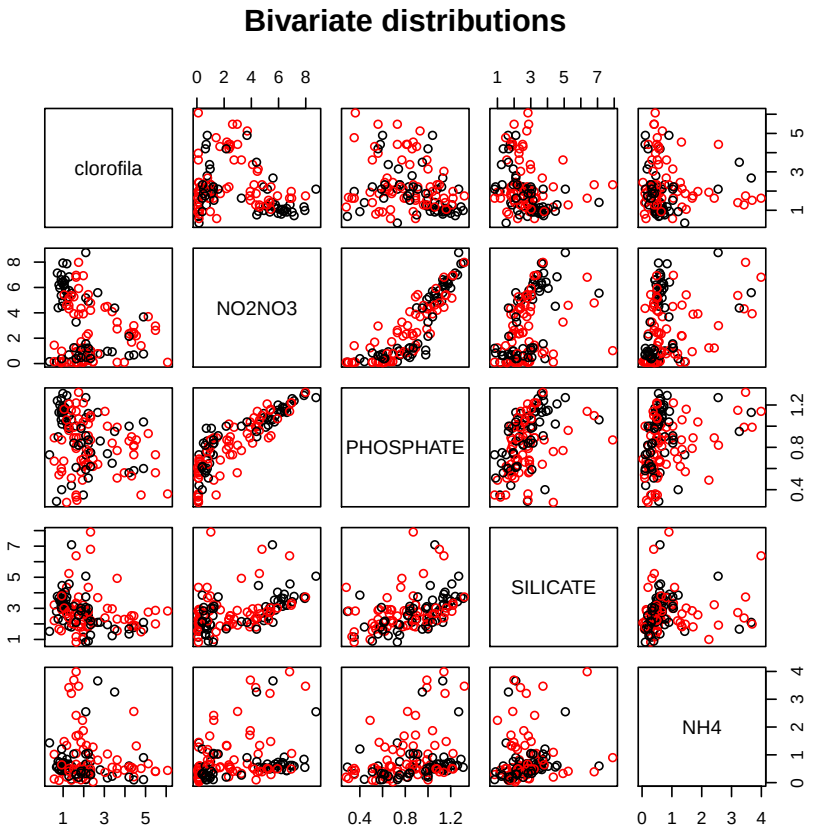

Figure 21: Bivariate distributions among variables. Red points refer to bivariate observations collected under the presence of nesting colonies while black points under abscence of seabirds.

4.2.4 Multidimensional scaling (MDS) and seabirds effect

The purpose of MDS is to represent data in a lower dimension while conserving the similarity that was present in the original distance matrix. Being y'_i and y'_j two observations in the r-dimensional space (in this case $r = 5$ and two objets might be two of the 125 samples) with dissimilarity value of d_{ij}, the purpose is to represent the original variables in a lower dimensional space (in this case $r = 2$) so that the dissimilarity computed in the new space d'_{ij} is as much possible as similar to the original d_{ij}.
Based on the euclidean distance matrix of the descriptors, Figure 22 represents a 2-dimensional version of the original data based on metric MDS. Furthermore, observations were splited into two groups according to presence and abscence of nesting colonies. It is not clear that there as a correlation between the (reduced) variables and the presence of nesting seabirds, which can be seen in the absence of a cluster among colors.

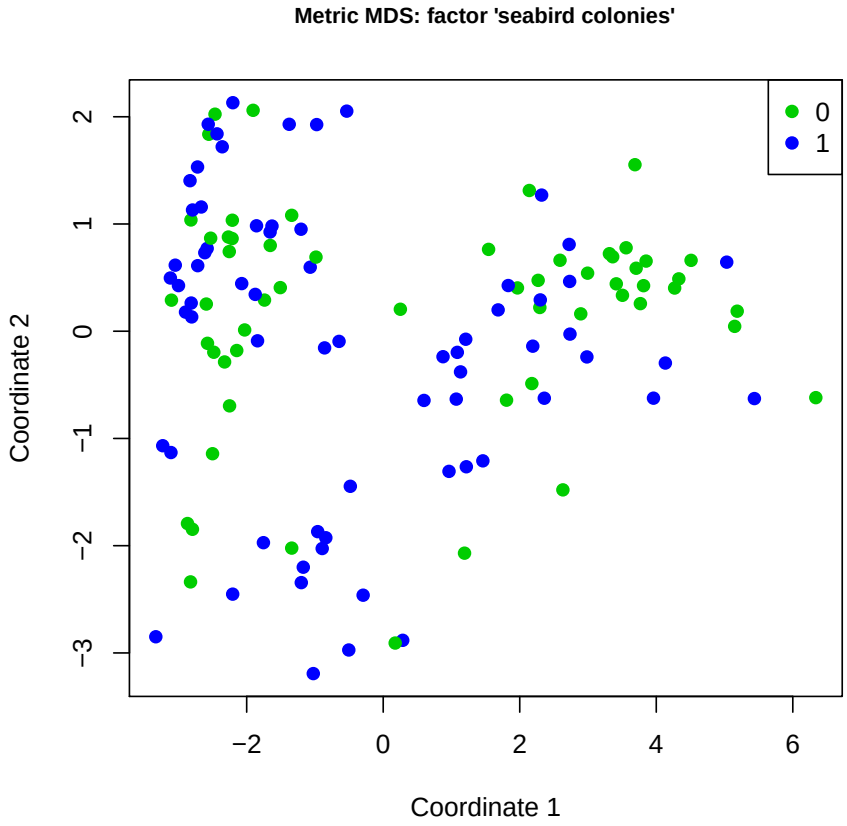

Figure 22: Mutidimensional scaling into 2 dimensions splitted into abscence (0) and presence (1) of seabird colonies.

Same analysis was made for a reduction in 3 dimensions following a metric MDS, as can be seen in Figure 23. The factor effect is not clearly distinguishable: again, levels seem assigned randomly.

Same results were obtained using a non-metric MDS (Figures not reported here)

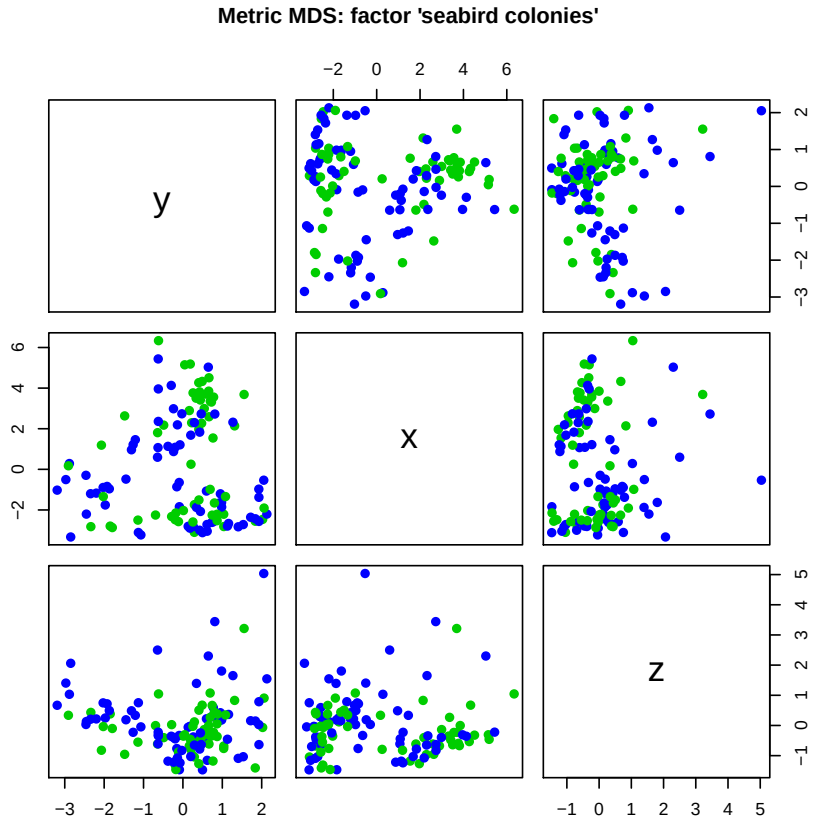

Figure 23: Mutidimensional scaling into 3 dimensions splitted into abscence (0) and presence (1) of seabird colonies.

4.2.5 Multidimensional scaling (MDS) and time effect

The association between the observations and the factor *period* was clear under a metric MDS (see Figure 24) as well as under a non-metric MDS (see Figure 25). The computed stress for the non-metric MDS was 8.99, which is considered between good and fair according to the Kruskal scale reported in Everitt, B.S., Dunn, G. (1991).

It is clearer than in the case of the factor seabird colonies' prescence, that 5 clusters are visually distinguishable. Nevertheless from both plots it seems that the observations recorded in September 2012 have more dispersion than the rest and this provoques some overlap beteween the samples of September 2012 and June 2012. In the non-metric MDS, the clusterization is more evident.

4.3 Model

The purpose was to test for significant differences in the multiresponse due to the presence or absence of seabird nesting colonies. To do that, a PERMANOVA procedure was implemented to a 2-way mixed model.

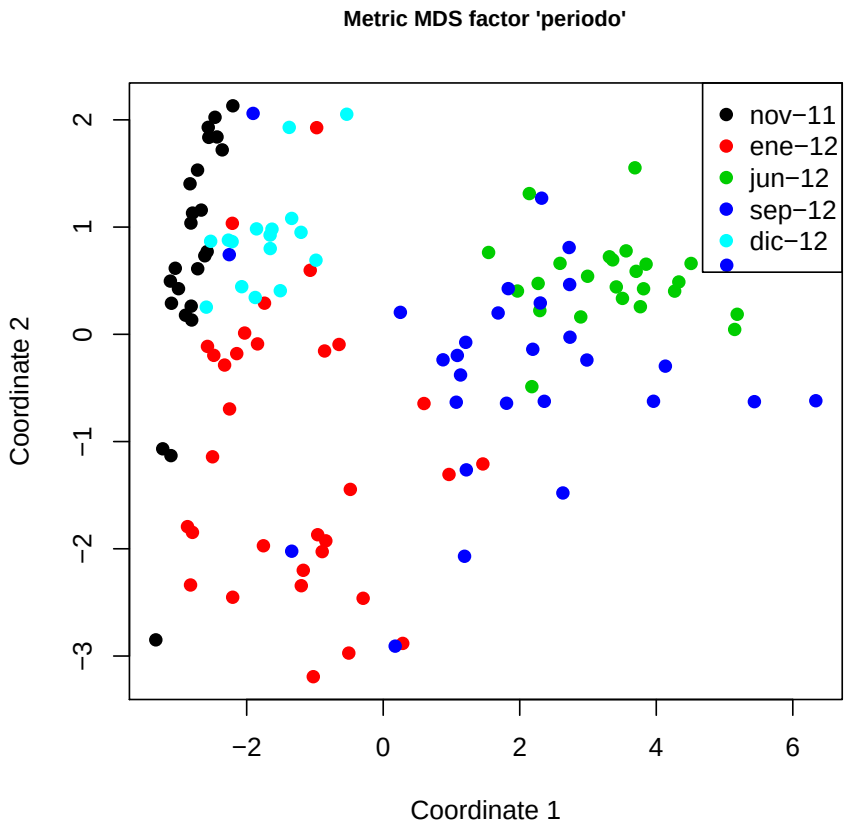

Figure 24: Metric mutidimensional scaling into 2 dimensions splitted into periods of measurement.

4.3.1 Factors

"Presence of seabird colonies" is the factor of interest and can be considered as fixed (from now on, factor A whose effect on the multiresponse is $\boldsymbol{\alpha}$). The factor has two levels ($a = 2$) according to the absence (level 0) and presence (level 1) of seabird colonies in the sampled locations (see Figure 19). Not all the above locations were sampled under both levels of the factor, resulting in some empty cells. The purpose in considering this factor is to reach some conclusions about the effects of the presence of seabird colonies over and above the effect already detected by enviromental factors represented by the factor *period*.

Period is considered a random factor with levels *Nov-11, Jan-12, Jun-12, Sep-12, Dec-12* (from now on, factor B, with $b = 5$ levels whose effect on the multireponse is $\boldsymbol{\beta}$). Month might also be considered as a fixed factor, but in this case the purpose is to estimate the variance component related to it. A factor may deserve the consideration of "random" when its levels are a sample from a wider population of possible levels, and this cannot be the case for a month, but what is being considered is the effect of climate uncertainity (Searle, 1992). Each level of the factor *period* (*Nov-11, Jan-12, Jun-12, Sep-12, Dec-12*) cannot be fixed as in the case of a fixed factor, because chosing a month like *november* doesn't mean that the researcher can replicate that condition in another

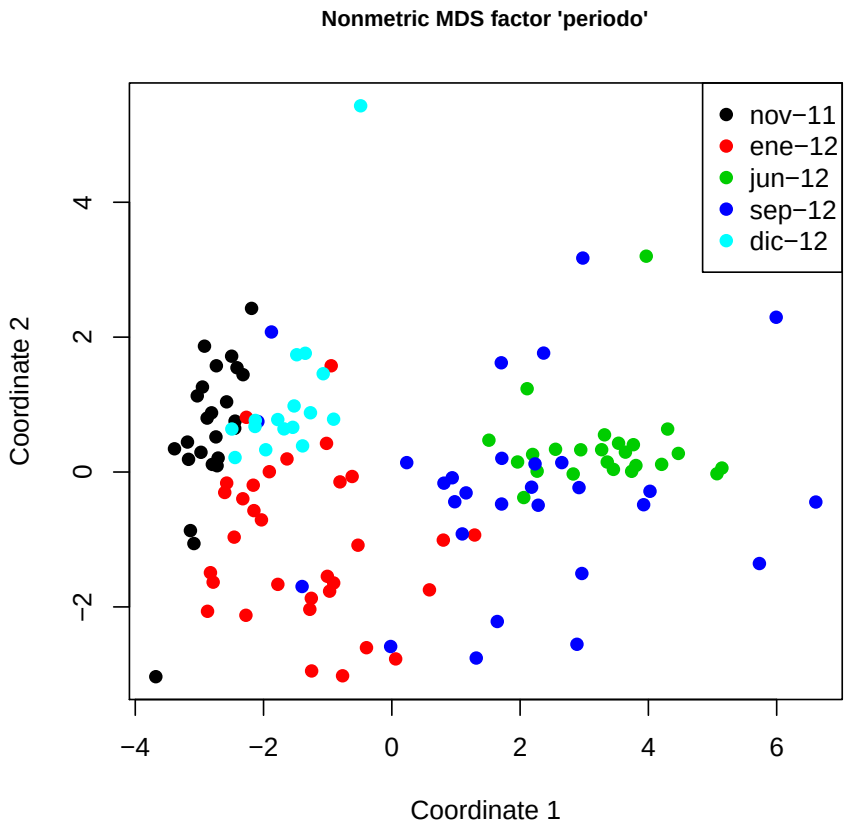

Figure 25: Non metric mutidimensional scaling into 2 dimensions splitted into periods of measurement.

experiment. Likely, all novembers will be more or less similar in terms of enviromental effects, but not the same, so their effect to the fixed factor will be random.

For a two factors design, one of which is fixed and the other is considered random there are three sources of variation to be tested: main fixed effect ($\boldsymbol{\alpha}$), the random effect ($\boldsymbol{\beta}$) and a interaction between them $A \times B$ whose effect on the multiresponse is $\boldsymbol{\lambda}$.

4.3.2 Model equation and assumptions

In an ANOVA, a model for the kth observation y_{ijk} where i denotes a level of factor A, j denotes a level of factor B and λ is the interaction effect, is represented by the following stochastic equation:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \lambda_{ij} + \epsilon_{ijk}$$
$$i = 1, 2, ..., a, \quad j = 1, 2, ..., b, \quad k = 1, 2, ..., n$$

The extension to the multivariate response $\mathbf{y}_{ijk}$, vector of p components or dependent variables is straightforward substituting the scalars by vectors:

$$\mathbf{y}_{ijk} = \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \boldsymbol{\beta}_j + \boldsymbol{\lambda}_{ij} + \boldsymbol{\epsilon}_{ijk}$$

where $\boldsymbol{\mu}$ is the overall mean and $\boldsymbol{\mu}$, $\boldsymbol{\alpha}_i$, $\boldsymbol{\beta}_j$ and $\boldsymbol{\lambda}_{ij}$ are all $p \times 1$ vectors. In addition to the model equation the following *zero-sum* restrictions hold:

$$\sum_{i=1}^{a} \boldsymbol{\alpha}_i = \sum_{j=1}^{b} \boldsymbol{\beta}_j = \sum_{i=1}^{a} \boldsymbol{\lambda}_{ij} = \sum_{j=1}^{b} \boldsymbol{\lambda}_{ij} = \mathbf{0}$$

The model equation has an error component which is the difference between the observation and its expected value,

$$\boldsymbol{\epsilon}_{ijk} = \mathbf{y}_{ijk} - E(\mathbf{y}_{ijk})$$

This error component includes the so called *experimental error* and any error related to it during the phase of collection, measurement and reporting. Errors are independent from each other and with distribution:

$$\boldsymbol{\epsilon}_{ijk} \sim f(\mathbf{0}, \Xi)$$

where f denote any probability distribution, not necessarily normal and Ξ is a matrix of variances. Covariances are zero since the errors are independent so $\Xi = \sigma_e^2 \boldsymbol{I}_n$, where n is the total number of samples.

Since B is a random factor, it characterizes the among unit variation, so the model assumes three sources of variations represented by $\boldsymbol{\beta}_j$, $\boldsymbol{\lambda}_{ij}$ and $\boldsymbol{\epsilon}_{ijk}$. The random effects follow a probability distribution, denoted by f_B since no paramatric assumption is made. $\boldsymbol{\beta}_j$ has $\mathbf{0}$ expectation and variance-covariance matrix $\boldsymbol{\Psi}$, whose dimension corresponds to the among-unit random effects in the model. It is assumed that $\boldsymbol{\Psi}$ has the same structure for all the different groups or levels, since there is no strong evidence that there are different conditions among the levels of the random factor. This is one assumption of the PERMANOVA procedure, even thoguh it is sensible to departures from homoscedasticity.

The inclusion of a random effect increases the variance of each $\mathbf{y}_{ijk}$ due to the random component. Furthermore, the interaction effect is also a random variable, so the total variance of the observations is:

$$V(\mathbf{y}_{ijk}) = \sigma_e^2 + \sigma_\beta^2 + \sigma_\lambda^2$$

4.3.3 Partitioning

As mentioned before, the PERMANOVA procedure is based on a partitioning of the distance matrix based on the multiresponse $\mathbf{y}_{ijk}$. Since any two vectors y'_{ijk} and $y'_{ijk'}$ are labeled according to the factor combination and the number of sample, the distance coefficient $d_{kk'}$ between them is also labeled according to the fact whether they are in the same factor combination or not.

Based on the partition, then pseudo F-ratios are built in order to test the three hypotheses of interest:

$H_{0(A)} : \boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2$ vs. $H_{1(A)} : \boldsymbol{\alpha}_1 \neq \boldsymbol{\alpha}_2$

$H_{0(B)} : \sigma^2_\beta = 0$ vs. $H_{1(B)} : \sigma^2_\beta > 0$

$H_{0(AB)} : \sigma^2_\lambda = 0$ vs. $H_{1(AB)} : \sigma^2_\lambda > 0$

The partition identifies, as is usual in ANOVA, the sources of variation, the sum of squares and the dregrees of freedom.

4.3.3.1 Sums of squares The sums of squares of the main effects are computed based on the sum of squares within each group and then doing the difference between the total sum of squares and the within sum of squares. The within sum of squares are the sums of the average interdistances withing each group. Being $d_{kk'}$ the distance between two objetcs i and j the following equations are applicable to the case of a balanced design in which each of the ab factor combinations has n replicates for a total of N multiresponse observations:

$$SS_{W(A)} = \frac{1}{bn} \sum_{k=1}^{N-1} \sum_{k'=k+1}^{N} d^2_{kk'} \phi^{(A)}_{kk'}$$
$$SS_{W(B)} = \frac{1}{an} \sum_{k=1}^{N-1} \sum_{k'=k+1}^{N} d^2_{kk'} \phi^{(B)}_{kk'}$$

$SS_{W(A)}$ and $SS_{W(B)}$ are the sums of squares computed on the distance matrix, so the sums of squared distances and

$$\phi^{(A)}_{kk'} = \begin{cases} 1 & if observations\ y'_k\ and\ y'_{k'}\ belong\ to\ the\ same\ group\ A \\ 0 & otherwise \end{cases}$$

and

$$\phi^{(B)}_{kk'} = \begin{cases} 1 & if observations\ y'_k\ and\ y'_{k'}\ belong\ to\ the\ same\ group\ B \\ 0 & otherwise \end{cases}$$

So the sum of squares for each of the main effects are:

$SS_A = SS_T - SS_{W(A)}$ and $SS_B = SS_T - SS_{W(B)}$

where $SS_T = \frac{1}{N} \sum_{k=1}^{N-1} \sum_{k'=k+1}^{N} d^2_{kk'}$

The residual sum of squares SS_R is computed following the same reasoning:

$$SS_R = \frac{1}{n}\sum_{k=1}^{N-1}\sum_{k'=k+1}^{N} d_{kk'}^2 \phi_{kk'}^{(AB)}$$

where,

$$\phi_{kk'}^{(AB)} = \begin{cases} 1 & if observations\ y'_k\ and\ y'_{k'}\ belong\ to\ the\ same\ group\ AB \\ 0 & otherwise \end{cases}$$

Indeed the residual sum of squares is computed in the same way as the error sum of squares are calculated in traditional ANOVA (from the within variability) but on the basis of interdistances.

The sum SS_{AB} of squares related to the interaction term is computed based on the difference:

$$SS_{AB} = SS_T - SS_A - SS_B - SS_R$$

4.3.3.2 Pseudo F-ratios Based on the information of the sum of squares, the construction of the statistics follows the sames rules as in traditional ANOVA.

A suitable test for the random effect would be provided by
$F_B = MS_B/MS_R$

where $MS_B = \dfrac{SS_B}{(b-1)}$ and $MS_R = \dfrac{SS_R}{(N-1)}$

For the fixed effect a suitable statistic must take account of the variability induced by the random factor, which is the variability due to the interaction term. The expected mean squares of the (random) interaction factor is an estimator of two components of variation (Montgomery, D.C. 2009):

$$E(MS_{AB}) = \sigma_e^2 + n\sigma_{AB}^2$$

while the expected mean square of the fixed factor estimates the above components of variation plus a term which is positive when the null hypothesis doesn't hold:

$$E(MS_A) = \sigma_e^2 + n\sigma_{AB}^2 + \frac{bn\sum_{i=1}^{a}\alpha_i^2}{(a-1)}$$

On the contrary, the expected mean square of the residual term is σ_e^2. So rejecting the null hypothesis by comparing MS_A with MS_R would leave open the question whether this is due to the factor's effect or to the (random) interaction variability. Therefore, a suitable test for the fixed effect would be provided by

$$F_A = MS_A/MS_{AB}$$

In a fixed design, testing for main effects would have sense after a test for interaction, but in the case of a mixed design, testing for a fixed effect has sense after a significant

interaction result because there is still the interest to check for main effects above the variability induced by the interaction term. A suitable statistic for the interaction term is provided by

$$F_{AB} = MS_{AB}/MS_R$$

as in the case of the fixed design, for the only difference between their expected mean squares is the component of variation due to the (random) interaction.

4.3.4 Partitioning for an unbalanced design

This section follows the content and conclusions in Searle (1971, 1987), Anderson, M. J.*et.al* (2008) and Maxwell, S.E. *et.al* (2004, Ch.7). As shown in table 12 there is unbalancedness and an empty cell in the data. This is not a case of planned unbalancedness or proportional ublancedness, which would make the procedures easier. Under this situation the above partitioning of sums of squares is not unique and since there is a relationship between the type of partitioning and the hypotheses being tested, care must be taken in order to chose one partition that tests a meaningfull hypothesis, where meaningfull in the context of this study is something very similar to the hypotheses above posed: that, in general terms, there is no difference between effects and that the variance components are all null. Following the advice of Anderson, M. J.*et.al* (2008), the so called Type III sum of squares will be given priority in this case of a 2-way mixed unbalanced model with an empty cell.

In general there are three things to consider about the types of sums of squares:

1. How is done the partitioning and which information is missed.
2. Which type of hypothesis is being tested by each type of partition.
3. What is the importance attributed to the sub-sample sizes. This is important because it is indeed the level of unbalancedness which determines the extent to which each partition's type will give comparable results. Sum of squares are reported sequentially:

Sums of squares can be structured in at least three ways (there is also a Type IV not included here):

1. Type I. It is a sequential report of sums of squares which add up the total sum of squares. It is the case, for example, of the function *ANOVA()* in the R language, which reports this way. The reported sums of squares are as follows:

    - SS_A from comparing a model with the A factor with a *only mean* μ model.

- $SS_B = SS_{(B|A)}$ from comparing the model $y = \mu + \alpha$ with the model with factor B, $y = \mu + \alpha + \beta$.

- $SS_{AB} = SS_{(AB|B,A)}$ from comparing the model $y = \mu + \alpha + \beta$ with the full model $y = \mu + \alpha + \beta + \lambda$.

The above sums of squares add up to SS_T so no information is left out. Since it is sequential, the sums of squares will be different for the two main effects (not for the interaction term) depending on the order in which are introduced in the model.

A Type I will take account of the weighted means, which means that it will consider each sub cell frequency of replicates. So it is important to note that might not be meaningfull when there is high unbalancedness or when there is some interest in considering each frequency.

2. Type II. The computation of sums of squares is not sequential. This type tests for each main effect after the other main effect:

   - $SS_A = SS_{(A|B)}$ so the model $y = \mu+\alpha+\beta$ is compared to the $y = \mu+\beta$ model.

   - $SS_B = SS_{(B|A)}$ comparing the model $y = \mu + \alpha + \beta$ to $y = \mu + \alpha$.

   - $SS_{AB} = SS_{(AB|A,B)}$ comparing the model $y = \mu+\alpha+\beta+\lambda$ with $y = \mu+\alpha+\beta$.

   Type II sums of squares will not add up to SS_T and are invariant with respect to the order in which main effects are fitted. This means that there is a loss of information: some sums of squares not being computed. Important is that this kind of procedure considers that there is no significant interaction effect. This is so because, after considering that the interaction rows $\times$ columns doesn't count then the columns (or rows) with more replicates will be given more weight in comparing means. So it is indeed a procedure which tests for differences in locations which depends on the relative size of the sub-samples.

3. Type III. It is another not sequential sums of squares which doesn't add up to SS_T. Each sum of squares of the main effects will be computed by comparing a model with the factor of interest with a model with the rest of the factors. For example, for factor A:

   $SS_A = SS_{(A|B,AB)}$ from comparing the model $y = \alpha$ with the model $y = \beta + \lambda$

The same for the factor B. The sum of squares for the interaction term is the same as in the Types I and II.
This partition is suitable when there is an interest in testing a hypothesis which doesn't depend on the cell frequencies because only the unweighted means will be compared. So it is suitable for the case of high unbalancedness or when there is interest in avoiding the the effects of unbalancedness. Unfortunatelly is the type with highest loss of information in the sums of squares due to the sums of squares that are missing in order to add up to SS_T

From the descriptions above I think the following elements can be considered in order to chose for a type:

- To take account of the importance given to the loss of information in Types II and III.

- To compare results under the partitioning's types of interest. In the present study I think it might be useful to compare Type I (which give full partition) with Type III wich is independent of the relative cell frequencies even though has loss of information.

- Important is to note that Types I and III test different hypotheses in the sense that Type I test means which are dependent from the cell frequencies while Type III does not.

- The sum of squares due to the interaction term is the same for all of the partition's types.

4.3.5 Variance components

4.3.5.1 Sources of variability The purpose of analysing the present data under a mixed model is to evaluate the fixed effect over and above the variability induced by the *period*'s effect. In such a model it is of interest to know the variance components, which are all the variance sources that add up the total variance of each observation. For a 2-way mixed model with interaction there are three sources of variance: the experimental error, the random factor and the interaction of the rabdom factor with the fixed one:

$$V(\mathbf{y}_{ijk}) = \sigma_e^2 + \sigma_\beta^2 + \sigma_\lambda^2$$

4.3.5.2 Computing components of variance Variance components are estimated by equating mean squares to their expected values and the construction of these expectations is explained in Searle (1992) and in Searle (1971, Chs. 9 and 10). The extension from the univariate case to the multivariate case is straight forward (Anderson, M. J. *et*

al., 2008).
The expected mean squares depend on the model since the values of the expectations depend on the randomness or not of each parameter. For balanced data, given the information for a 2-way crossed design with interaction:

- a: number of levels in factor A
- b: number of levels in factor B
- n: number of (balanced) replicates in each combination

the expected mean squares are:

1. $E(MS_A) = \sigma_e^2 + n\sigma_\lambda^2 + \dfrac{bn\sum_{i=1}^{a}\alpha_i^2}{(a-1)}$
2. $E(MS_B) = \sigma_e^2 + an\sigma_\beta^2$
3. $E(MS_{AB}) = \sigma_e^2 + n\sigma_\lambda^2$
4. $E(MS_R) = \sigma_e^2$

Then the variance components estimattions are otained equating the above values to the computed mean squares according to the sums of squares table.

4.3.5.3 Components of variance for unbalanced data There are two main differences in the case of unbalancedness when compare to the previous case:

1. For each of the types of sums of squares, the means of squares will be also different, expect for the case of interaction. This leads to different equations when equating expected mean squares to computed mean squares. So the estimation of variance components will depend on which type of partition is chosen.

2. The coefficients that are used to combined the variance components, an, bn, and n are no longer whole numbers. These arise from different methods to compute the variance components, the so called *Hendersons' methods* and other procedures explained in chapter 10 of Searle (1971) and summarizes in a formulae's chapter in Searle (1971, Ch.11)

4.4 Results

4.4.1 Implementation

All the analyses were performed with PERMANOVA+ (Gorley *et al.*), which is an-add on package of Primer 6, and R (R Development Core Team (2014). PERMANOVA+ was the software employed in all the analyses regarding the test procedures of Section 3.4. In R, the PERMANOVA procedure can be implemented in a limited way (only for a cross factorial designs, even nested but without random effects) by the function *adonis* of the package *Vegan* (Oksanen, J. *et al.* 2001). The function reports the Type I partition's sum of squares.

4.4.2 Mixed model with two factors

As stated before, the main purpose was to test the effect of seabird nesting colonies (actually their guano) over and above the variability induced by *period*, since time variation is already known by researchers. An analysis of variance based on an euclidean matrix distance of the variables $\mathbf{y}_{ijk}$ was performed using PERMANOVA+ for the following model:

1. Factors:
   - Seabird Colonies (GR)
     - Levels: 0 (abcense of colonies); 1 (precense)
     - Type: fixed effect
   - *Period* (T)
     - Levels: Nov-11; Jan-12; Jun-12; Sep-12; Dec-12
     - Type: random effect
2. Model equation:
   $\mathbf{y}_{ijk} = \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \boldsymbol{\beta}_j + \boldsymbol{\lambda}_{ij} + \boldsymbol{\epsilon}_{ijk}$
3. Variance components:
   - σ_e^2: residuals
   - σ_T^2: random factor *period* (T)
   - σ_{int}^2: interaction of factors

4.4.2.1 Results based on Type III sums of squares Results based on a Type III partition are reported in Table 15. The null hypothesis in this test was that *seabird colonies* had no effect on the nutrients and chlorophyll values. The permuted P-value was 0.175, based on 9999 permutations, 9680 of them unique (only a few were repeated orderings). This means that we could't find a strong evidence in favour of the hypothesis that the presence of *seabird colonies* affects the levels of nutrients. The above P-value is the probability of observing an extreme value as the observed, which is $F_{GR} = 2.585$, if

PARAMETERS AND ANALYSIS OF VARIANCE						
Parameters						
Model	2-factors mixed					
Factors	GR Bird colonies (fixed) T *Period* (random)					
Sums of squares type	III					
Permutation method	Unrestricted permutation of raw data					
Number of permutations	9999					
Distance	Euclidean					
Analysis of variance						
Source	df	SS	MS	Pseudo-F	Significance	U.Per.
GR	1	22.65	22.65	2.585	0.175	9680
T	4	590.72	147.68	41.919	0.0001	9932
Interaction	3	27.43	9.143	2.595	0.022	9927
Residuals	116	408.66	3.523			
Total	124	1236.1				
Variance components estimation						
σ^2_T	7.295					
σ^2_{int}	0.553					
σ^2_e	3.523					

Table 15: PERMANOVA results for a 2-way mixed model based on a Type III partition.

H_0 holds and this was higher than the level of significance α when it is set at 0.05.

In the context of the mixed models, as stated before, a proper pseudo F-ratio for the fixed effect is computed by comparing the mean square of the fixed effect with the mean square of the interaction term, since comparing it with the mean square of residuals would have left unexplained the random effect of the interaction between the fixed effect and the random one. Note that if this had been a fixed design, the pseudo F-ratio would have been approximately $22.65/3.523 = 6.429$ and likely the permuted P-value would have been less than 0.175 because there would have been less permuted F-ratios greater than 6.429.

Both the random factor (T) and the random interaction term ($Interaction$) were significants when compared to $\alpha = 0.05$, with permuted P-values of 0.0001 for T and 0.022 for $Interaction$). This means that there was significant variability among periods and significant interaction. The strong variability among periods was expected, but the purpose was to test the fixed effect conditioned on this variability. Indeed this variability was already observed in the mutidimensional scaling plot per period of figure 24 and figure 25.

The estimation of variance components allowed to split the variance of $\mathbf{y}_{ijk}$ according to the sources of variability. The residual variability was found equal to 3.523 and the variability due to *period* added 7.295 to the whole variability which was, considering also the interaction term, 11.371. So the random factor explained 64.15% of the whole variability.

The decision based on the interaction's term result doesn't work as in the case of the fixed factor. In a 2-way fixed design, the significance on the interaction term makes the results about the main effects not meaningfull, but it is not the case in a mixed model, since there is still interest about the fixed effect over the noise induced by the random components.

Type III didn't add up to SS_T, indeed all the sources of variation in Table 15 added up to 1049.46, leaving 186.64 of sum of squares not explained by any component. The advantange of this type of partition is that the tested null hypothesis of equality of means doesn't consider each subsample frequency, which might be of interest in the case of a very unbalanced design (see Table 12). It is also recomended in the case of a design with empty cells, as in the present case (Anderson, M.J. *et al. 2008*)

As for the whole model, the fixed effect factor explained just a small part of the total source of variation equal to 1236.1, since it was only 22.65. The estimated model esplained 51.8% of the total variability considering that some variability was left out due to the kind of partition. It means that there is a relevant part of the variability still in the error. Perhaps the inclusion of other terms, especially the spatial factor (*location*), could have improved the explanation capacity of the model, but as mentioned before, for such a design there was not enough information (many empty cells with high unbalancedness, see Table 11) so it is not clear how meaningfull results based on this information might have been.

4.4.2.2 Pairwise comparisons based on Type III Even though the permuted P-value was found greater than 0.05, it can be considered as a meaningful trend of the effect above the noise of the random factor, so it might be interesting to build permutational *t-statistics* to compare the fixed effect's levels within each of the levels of the random factor, having in mind that these are not fixed but a sample from a population of levels.

The PERMANOVA procedure considers such a test as a one-way with two levels, since both levels of the fixed factor are compared within each level of the random factor. Following the analogue to traditional ANOVA, the procedure is based on a pseudo t statistic for the multivariate case, which is the square root of the pseudo F-ratio. Again the multivariate information is reduced in a distance matrix (actually it is the distance matrix of the whole analysis) and the distance "between" and "within" distances are compared: "within" if both observations belong to one level (0) or the other (1) of the fixed effect *seabird colonies*, and "between" if one observation is in one level and the other in another level.

Such tests where computed using an euclidean distance and through an unrestricted permutation of raw data. Reported values in Table 16 are based on a Type III partition of sum of squares. Lower triangular matrices show the average distances within groups (subindices $(0,0)$ and $(1,1)$) and between groups (subindex $(0,1)$). Furthermore, since during June 2012 there were no replicates to test for the presence of nesting colonies, there is no chance to test for a difference within this period. As can be seen, the difference in the fixed factor was strongest in January 2012 (Jan-12) and September 2012 (Sep-12) with permuted P-values (based on 9999 permutations) 0.0001 (reported as 0.000) and 0.037 respectively.

PAIRWISE TESTS WITHING LEVELS OF THE RANDOM FACTOR			
Period	t value	Permuted P-values	Average distance
Nov-11	1.234	0.197	$\hat{d}_{i,j} = \begin{pmatrix} 1.627_{0,0} & \\ 1.812_{0,1} & 1.919_{1,1} \end{pmatrix}$
Jan-12	3.353	0.000	$\hat{d}_{i,j} = \begin{pmatrix} 1.609_{0,0} & \\ 2.884_{0,1} & 2.471_{1,1} \end{pmatrix}$
Jun-12	NA	NA	NA
Sep-12	1.696	0.037	$\hat{d}_{i,j} = \begin{pmatrix} 4.474_{0,0} & \\ 4.046_{0,1} & 2.973_{1,1} \end{pmatrix}$
Dec-12	1.579	0.056	$\hat{d}_{i,j} = \begin{pmatrix} 1.124_{0,0} & \\ 1.911_{0,1} & 2.346_{1,1} \end{pmatrix}$

Table 16: Pairwise comparisons of seabirds effects within each period.

4.4.2.3 Results based on Type I sums of squares Type I sums of squares allowed for all the sources of variation to be included in the analysis. Sums of squares reported in Table 17 added indeed up to SS_T, so no variation was left out.

The permuted P-value for the factor of interest was found equal to 0.447 which is higher than what reported based on Type III partitioning. In this case, nevertheless, the hypothesis of equality of means considers the frequency of samples in each subcell, while Type III is based on unweighted means. Based on Type I, the fixed effect couldn't be considered significant above the noise induced by the *period*.

As in the previous case, there was a significant variability among periods with a permuted P-value of 0.0001. Also the random interaction factor was found significant and this was expected since for each type of partition of sums of squares the interaction term's sum of squares is the same. The variability due to *period* was 67.7% of the total variance of $\mathbf{y}_{ijk} = 8.146 + 0.553 + 3.523$ (see Table 17).

PARAMETERS AND ANALYSIS OF VARIANCE						
Parameters						
Model	2-factors mixed					
Factors	GR Bird colonies (fixed) T *Period* (random)					
Sums of squares type	I					
Permutation method	Unrestricted permutation of raw data					
Number of permutations	9999					
Distance	Euclidean					
Analysis of variance						
Source	df	SS	MS	Pseudo-F	Significance	U.Per.
GR	1	44.95	44.95	0.603	0.447	9806
T	4	755.05	188.76	53.581	0.0001	9926
Interaction	3	27.43	9.143	2.595	0.020	9933
Residuals	116	408.66	3.523			
Total	124	1236.1				
Variance components estimation						
σ^2_T	8.146					
σ^2_{int}	0.553					
σ^2_e	3.523					

Table 17: PERMANOVA results for a 2-way mixed model based on a Type I partition.

4.4.2.4 Comparisons based on types of permutations As mentioned in Section 2, there are at least three types of permutations: unrestricted permutation of the raw data, permutation of residuals from a reduced model and permutation of residuals from a full model. As can be seen from Table 18 (from partitions of sums of squares based on Type III) the trend among all types of permutations was the same: the permuted P-value for the fixed effect was moderate and for the random factor *period* was significant when compared to $\alpha = 0.05$. For the interaction there seemed to be a trend toward more significance when permuting residuals. Anderson, M.J. *et al.* (2003) performed a simulation study in order to compare several permutation methods in terms of power and proportion of rejections of the true null hypothesis. The aim of the study was to

compare properties for data simulated under 4 different distribution functions -normal, uniform, lognormal and exponential of parameter 1, whose values were then cubed-. The last parametric model was used to generate variates radically non-normal. The test for the fixed effect in a mixed model was found to perform relatively poorly, with a proportion of rejections of the true null hypothesis reaching 7% in the case of exponential random variates both, in the case of permutation of raw data and residuals, but not for the remaining distributions (see Table III in Anderson, M.J. *et al.*, 2003). As mentioned before, the test for interaction based on residuals approximates asymptotically the exact test. Worth to note that all these simulations were based on balanced data.

	Permuted P-values		
Type of permutation	Fixed effect	Random effect	Interaction
Unrestricted permutation of raw data	0.175	0.000	0.022
Permutation of residuals under a reduced model	0.125	0.000	0.006
Permutation of residuals under a full model	0.100	0.000	0.007

Table 18: Permuted P-values' comparison among different types of permutations.

4.4.2.5 Permutational analysis of multivariate dispersions It has been stated before that the PERMANOVA procedure is sensible to different dispersions in the multivariate distributions among groups. Indeed a significant P-value for a term might be due to a true difference in locations (the real target of the PERMANOVA procedure) or due to different dispersions among the groups of interests.

4.4.2.5.1 A test on dispersions based on distances Following the method of analysis of variance based on distances, let $\mathbf{y}_{ij}$ be a vector of p components (variables) where i denotes one group or treatment and j denotes the observation. Let $\bar{\mathbf{y}}_i$ be the vector of p means in the i^{th} group. Then,

$$w_{ij} = d_E(\mathbf{y}_{ij} - \bar{\mathbf{y}}_i) = \sqrt{\sum_{k=1}^{p}(x_{ijk} - \bar{x}_{ik})}$$

is the euclidean distance among all the p components between all the observations belonging to the i^{th} group.

For each group, an average distance to centroid is computed and an ANOVA is performed to compare the w_{ij}s among each of the i^{th}s groups. If a cloud of points has more dispersion than the other then its distances will be, on average, larger, being likely a significant difference among groups' distances w_{ij} (see Figure 26 for two dimensions). A pseudo F-statistic is computed and a permutation P-value is estimated comparing the F observed to a permuted distribution in which the statistic is calculated upon each permutation of data.

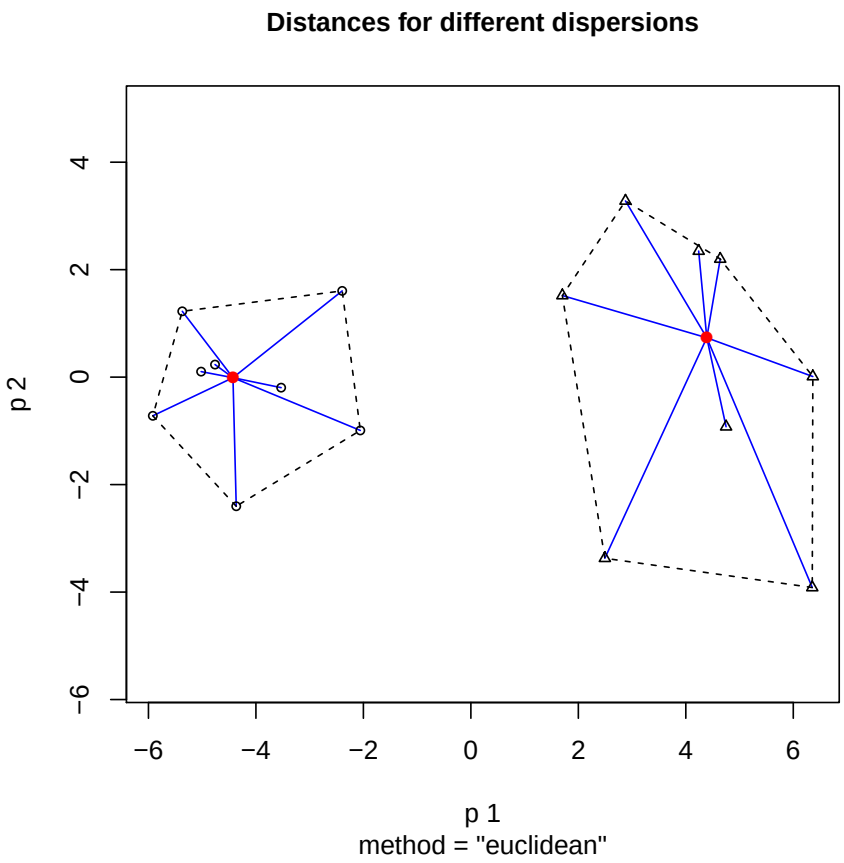

Figure 26: Difference in distances to the centroid for two groups of points.

The PERMANOVA procedure is flexible for being used on several types of dissimilarities. Indeed, it might be of interest not to use an Euclidean distance to represent resemblance. In cases like the Bray-Curtis resemblance measure, the centroid cannot be computed on the arithmetic sample mean. The procedure is then to calculate principal coordinates from the distance matrix and then to compute the centroids based on the sample means as well as the w_{ij}s values.

4.4.2.5.2 Test on dispersions: results First, a test was computed for the factor of interest *seabird colonies*, for which the comparison was between two mean distances. As can be seem in Figure 27 there seemed to be no serious difference in dispersions. In the right part of Figure 27, distances are represented as blue rows, and the two red points, scarcely visible, are the mean centroids. It can be seen that the distances are not very different in the two groups. Indeed, the difference in the location between the two groups (measured by the median) -in the boxplot- seemed non relevant. The multivariate response in Figure 27 was based on the principal coordinates of the 5 components of the response data.

Average distances to centroides in each group where found very similar (see Table 19) and the permuted P-value was 0.154, which can be interpreted as a trend toward different

dispersions, but they were not significant based on the comparison with the usual $\alpha = 0.05$.

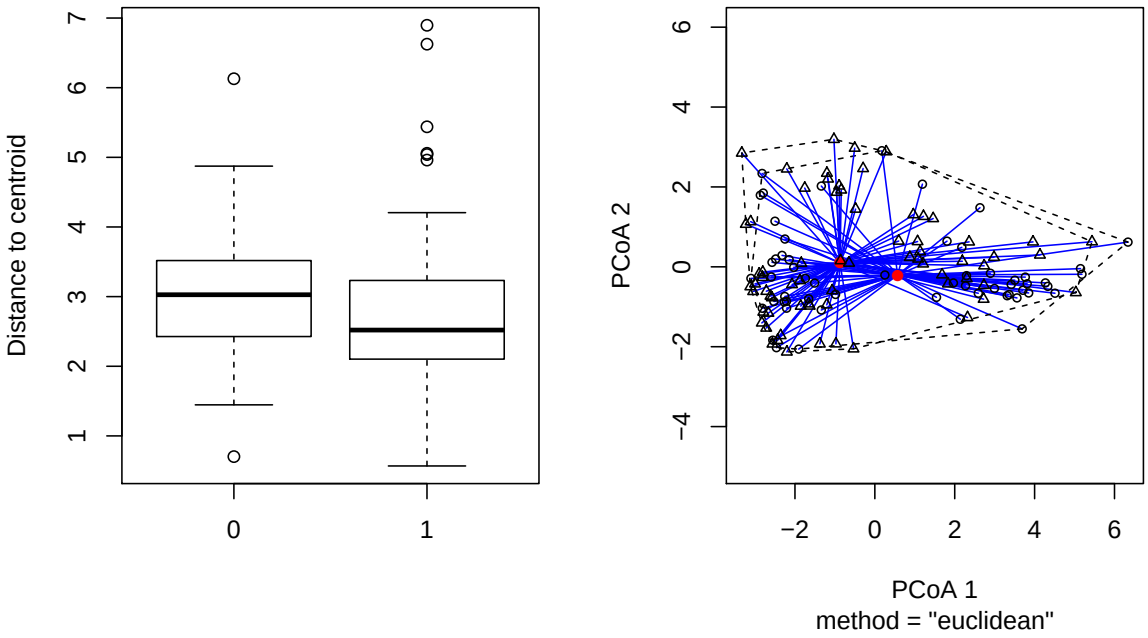

Figure 27: Dispersions in each level of the fixed factor represented in 2 dimensions through the principal coordinates.

Group level	Average distance
0 (absence)	3.069
1 (presence)	2.796
F-statistic observed	2.336
Perm P-value	0.154
Permutations	999

Table 19: Test on dispersions for the factor of interest *seabird colonies*.

A second test was performed for the levels of the factor *period* (5 levels). In this case, as can be seen from Figure 28 there seemed to be a clearer difference in the distances to centroids. The patterns of variability among the periods were very different: *Nov-11* had many distance values outside the interquartile range and *Dec-12* had one value which can be consider an outlier. It can be seen that the central location of distances, represented by the median, were also different, which reinforces the hypothesis of different mean distances or different dispersions.

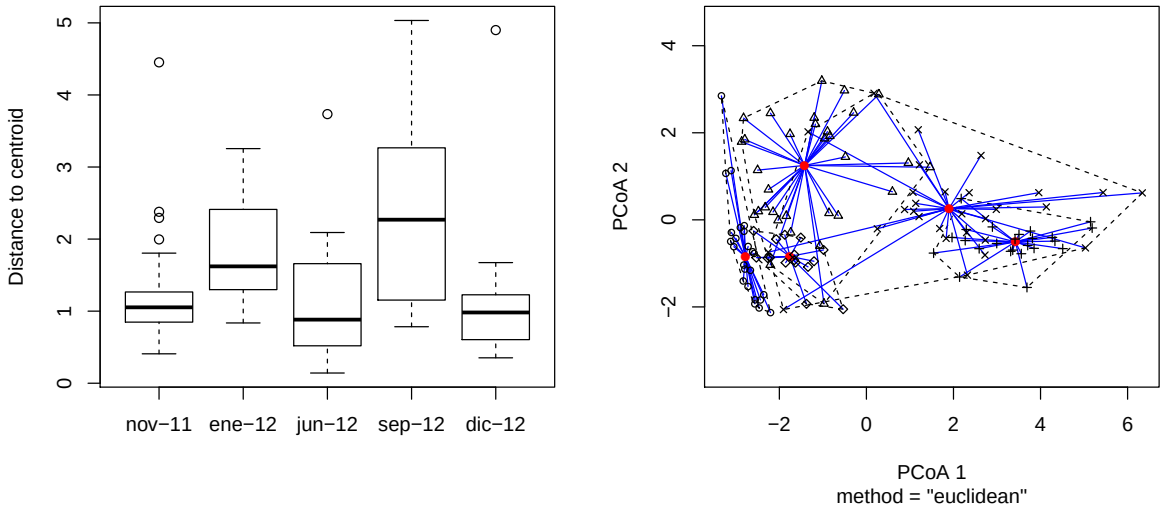

Figure 28: Dispersions in each level of the random factor represented in 2 dimensions through the principal coordinates.

The difference in dispersions was found significant with a permuted P-Value of 0.0001 (see Table 20) and this was indeed expected since for a random factor, the relevant test in a 2-way mixed design is precisely related to differences in variability instead of locations.

Group level	Average distance
Nov-11	1.286
Jan-12	1.821
Jun-12	1.114
Sep-12	2.459
Dec-12	1.224
F-statistic observed	10.292
Perm P-value	0.0001
Permutations	999

Table 20: Test on dispersions for the random factor.

4.4.2.6 Introduction of a covariate Researchers collected data about the distance of the point of observation to coast. This was considered, *a priori*, a usefull source of enviromental information and, depending on the distance, an indicator of wave exposure. This last covariate was, for example, considered as a factor in Kolb, G.S. *et al.* (2010).

Distance to coast is a continuous variable measured in meters and, as can be seen from the frequency plot (Figure 29), a very skewed random variable. As can be seen from the bivariate plots in Figure 30 nutrients and chlorophyll seemed to be not affected by the distance to coast. A PERMANOVA procedure was performed to confirm this impression considering the fixed factor *seabird colonies*, the random factor *period* and the covariate *distance to coast*. As can be seen from Table 21 there was no relevant improvement in terms of sums of squares, since there still were 408.30 out of 1236.1 sums of squares in the residuals, and the covariate *Distace to coast* only represented 4.5 units of sums of squares. Furthermore the permuted P-value due to the covariate was 0.9054, which indicated that we couldn't find evidence that nutrients' values were affected by the "Distace to coast". The partitioning was of Type I, so the sums of squares of the fixed, random and interaction terms were computed after fitting the covariate (Type I is a sequential partitioning). This analysis also confirmed the trends observed in the previous sections regarding the fixed effect and the random effect.

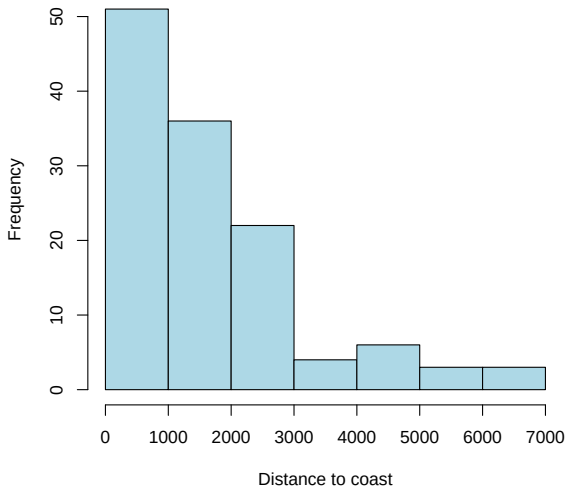

Figure 29: Frequency of the covariate *Distance to coast*.

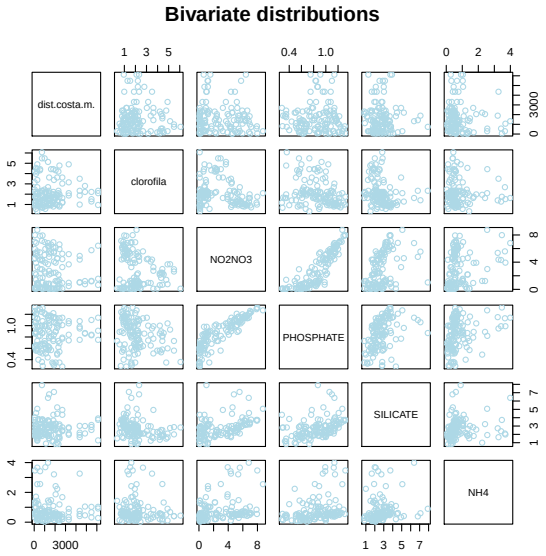

Figure 30: Bivariates sample distributions between *Distance to coast* and nutrients and chlorophyll. As it is clear from the bivariate plots of the first column, there is no clear pattern of relationship between the distance and nutrients.

4.4.2.7 *Location* as random factor As shown in Table 11, there was not enough information to allow for a mixed 3-way analysis involving the spatial (*location*), the *period* and the fixed factor *seabird colonies* together. Nevertheless since the spatial aspect was of interest *a priori*, a PERMANOVA procedure was performed considering a 2-way mixed model by treating *location* as a random term, since it was of interest the fixed effect above the spatial variability.

PARAMETERS AND ANALYSIS OF VARIANCE						
Parameters						
Model	2-factors mixed					
Factors	GR Bird colonies (fixed) T *Period* (random) Dist Distance to coast (Covariate)					
Sums of squares type	I					
Permutation method	Permutation of residuals under a reduced model					
Number of permutations	9999					
Distance	Euclidean					
Analysis of variance						
Source	df	SS	MS	Pseudo-F	Significance	U.Per.
Dist	1	4.50	4.50	0.243	0.9054	9949
GR	1	54.74	54.74	0.762	0.4359	9816
T	4	741.16	185.29	52.187	0.0001	9934
Interaction	3	27.39	9.13	2.571	0.051	9943
Residuals	115	408.30	3.55			
Total	124	1236.1				
Variance components estimation						
σ^2_T	8.112					
σ^2_{int}	0.557					
σ^2_e	3.551					

Table 21: PERMANOVA results for a 2-way mixed model with *Distance to coast* as covariate based on a Type I partition.

As can be seen in Table 22, the sums of squares partitioning changed drastically when compared to the 2-way mixed model with *period* as random factor. The permuted P-value was found equal to 0.0071, which gave evidence in favour of the hypothesis that the presence or absence of seabird colonies had an effect on the levels of nutrients. But this was clearly due to the fact that the factor of interest is absorbing the variability due to the *period*.

Not only this model still had 888.16 of sums of squares among the residuals, which was a higher value compared to the 2-way mixed model with *period* as random factor. Furthermore, this model failed to give an answer to the main research question, whether there was a relevant effect due to *seabird colonies* that can be distinguished from the *period* cycle.

Note also that, in this model, the highest variance component was provided by the

residuals (7.723) and not by the random effect (3.323). The variance estimated for the interaction was negative, which is an unusual value for a variance. This is indeed a chance in variance components estimations. Searle (1971) suggested several choices to deal with negative estimates of variance components:

- Accepting the estimate and considering it as evidence that the true value of variance is zero (evidence in favor of the null hypothesis);
- Using 0 as estimate, so replacing the value;
- Excluding the factor of interest;
- Considering it as evidence of a wrong model;
- Using other methods, such as Bayes procedures.

I think that, considering that it is a negative but small value, it can be regarded as zero.

4.4.3 Further discussions

From the previous results, it is clear that the main problem in achieving the purpose of the study was the lack of information for a complete 3-way design, which might have considered the fixed factor of interest, *seabird colonies*, and the random terms *location* and *period* in a single test. This was clearly recognizable in the grid of Table 11, in which there were many empty cells and lack of the so called *geometrical connectedness* (Searle, 1987).

In certain studies, such as classification studies, unbalancedness is very likely. Also in clinical trials, where some patients could drop a given treatment. In these cases one might say that unbalancedness arise from true differences in population sizes. A situation which might lead to empty cells is *selection bias*, whether the treatment conditions (in the present case the presence of nesting colonies, the period of measurement and the location) are or not differencially responsible for the unbalancedness and/or empty cells (Maxwell, S.E. *et al.*, 2004). Selection biases affect the meaningfullness of the results for an unbalanced design because the true effect will be confounded with preexisting differences. In the present design some locations, T_1 and T_6, were expressively chosen because of their absence of nesting colonies. As can be seen from Table 13, there were 12 and 25 replicates in coincidence of absence of seabird colinies. In this case the decision on the random effect (choosing levels T_1 and T_6) led to some cells empty. Selection biases leading to empty cells will affect the estimation of the interaction terms (Searle, 1087). In the same way, the month June 2012 adds variability only to the level *no presence of nesting colonies* since during this month there are no seabirds in the area.

What led to this lack of complete information were physical conditions, difficulties when trying to reach some points of observations in certain periods. This resulted, as was reported in Tables 15, 17 and 22 in difficulties in order to reduce the residual sums

PARAMETERS AND ANALYSIS OF VARIANCE						
Parameters						
Model	2-factors mixed					
Factors	GR Bird colonies (fixed) L Location (random)					
Sums of squares type	III					
Permutation method	Unrestricted permutation of raw data					
Number of permutations	9999					
Distance	Euclidean					
Analysis of variance						
Source	df	SS	MS	Pseudo-F	Significance	U.Per.
GR	1	254.64	254.64	71.564	0.0071	9731
L	5	281.20	56.24	7.282	0.0001	9930
Interaction	3	10.467	3.56	0.460	0.8633	9943
Residuals	115	888.16	7.72			
Total	124	1236.1				
Variance components estimation						
σ^2_L	3.323					
σ^2_{int}	-0.541					
σ^2_e	7.723					

Table 22: PERMANOVA results for a 2-way mixed model based on a Type III partition with *location* considered as a random effect.

of squares and in giving the random factors, one at a time, relevant weight as variance components.

5 Conclusions

The purpose of the present study was both to empirically analyse the performance of the PERMANOVA test under certain conditions that were found in the data collected by researchers at Puerto Madryn in Argentina and to answer to the research question, *Does the presence of seabird colonies affect the values of nutrients and chlorophyll?* So in order to answer it we had to consider the main patterns that we found during the simulations as well as the limitations of the computational analysis.

The collected data at Puerto Madryn, due to physical limitations, didn't obey to a well designed procedure, leading to high unbalancedness and empty cells. Indeed, two *a priori* important sources of variation, *location* and *period* couldn't be simultaneously examined with the levels of the effect of interest (*seabird colonies*) in a 3-way mixed design due to a high number of missing replicates, which led to a layout as given in Table 11. According to Searle (1987), significant lack of cennectedness between the cells in a grid would make the results non meaningfull.

The results on real data were therefore based on the simultaneous analysis of the factor of interest, *seabird colonies*, and the random factor *period* in a crossed mixed design, with a number of replicates for each factor level combination as given in Table 12. The found P-value (0.175, see Table 15) couldn't be considered a strong evidence in favour of an effect of *seabird colonies* on the values of nutrients and chlorophyll. The PERMANOVA procedure comfirmed what was evident during the exploratory analysis, that the *period* is an important source of noise and in all cases explained the highest portion of the sums of squares.

Such a crossed design, with $N > p$ (in the real case, $N = 125$ and $p = 5$), with an unequal number of replicates and missing cells, was considered in the simulation study. As can bee seen in Table 23, there were no replicates for the second level of *seabird colonies* (1, presence) and the level *Jun-12* of *period*. This unbalancedness case was defined during the simulations as *U (with 1 empty cell)*.

	Nov-11	Jan-12	Jun-12	Sep-12	Dec-12	Totals
0 (absence)	4	12	24	9	8	$N_{1.} = 58$
1 (presence)	20	20	0	20	8	$N_{2.} = 68$

Table 23: Number of replicates available for the 2-way crossed design *period* × *seabird colonies* with subtotals. The sub-sample $N_{a=2,b=3}$ has no replicates.

Table 24 summarizes the performance of the PERMANOVA test when the simulations where run in scenarios similar to those encountered in the real data. Of course, neither the true distribution of the real data, nor the veracity of the null hypothesis could be accerted for the real case, since these are unknown. So in order to interpret the real results in the light of the simulations results, we considered the scenarios were $N = 120$, which is closed to the real $N = 125$, the N/p ratio was greater than 1 (in the real case $N/p = 25$)

and the employed distance was, as in the real data, Euclidean. As could be seen from the simulations results, in this context, the PERMANOVA test performed closed to the nominal level of 0.05 when data was homoscedastic (the cases of $k = 1$ an $v = 1$) for both factors A and B, both when the simulated distributions were multivariate normal and negative binomial (real rejections can be traced in Tables 25 and 26, which were delimited to facilitate the identification of the P-values).

		Performance of PERMANOVA			
	N	$k = 1$	$k = 3$	$k = 10$	$k = 20$
$N > p$, Euclidean distance, MVN distribution and *U (with 1 empty cell)*	120	closed (closed)	conservative (liberal)	conservative (liberal)	conservative (liberal)
		$v = 1$	$v = 0.05$	$v = 0.02$	$v = 0.01$
$N > p$, Euclidean distance, Nbin distribution and *U (with 1 empty cell)*	120	closed (closed)	conservative (closed)	closed (liberal)	closed (liberal)

Table 24: Performance of the PERMANOVA test under simulation scenarios similar to those found for the real case. The top label refers to the A factor and the bottom label, in parenthesis, refers to the B factor.

The closest scenario to the real case was when $k = 1$ for two reasons: first, k refered to the multivariate normal distribution and eventhough the real distribution is unknown, at least can be considered as sampled from a continuous distribution. Second, $k = 1$ refers to a homoscedastic situation in relation to the factor A or *seabird colonies* in the real case. As could be confirmed by a distance-based test on dispersions in 4.4.2.5.2, non strong evidence of heteroscedasticity could be found, with a P-value of 0.154 (see Table 19).

Some limitations to the above conclusions are worth to be mentioned and constitute a set of conditions upon which to develope further the present study. First, even though the real case and the simulations involved crossed designs, the first one had $b = 5$ levels belonging to the second factor while the simulated case had $b = 3$, in order to limit the computing time. Second, simulations considered a fixed design while the real case was analysed under a mixed model. As already mentioned, function *adonis* of *Vegan* package is of direct use for fixed factors designs. Including mixed or pure random designs would require further computational modifications of the code that were beyond the purpose of this work. This is related to the third limitation. Even though the *Vegan* package is widely used and well referenced among the R community, for us was not like having an R code of direct control. A very interesting future line of applied study would be, indeed, to build an on-purpose code in R allowing for simulations under several complex designs.

6 Bibliography

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral. Ecol.*, 26, 32 - 46.

Anderson, M. J. and Robinson, J. (2001). Permutation tests for linear models. *Austral. & New Zealand J. Statist.*, 43, 75 - 88.

Anderson, M. J. and Robinson, J. (2003). Generalised discriminant analysis based on distances. *Austral. & New Zealand J. Statist.*, 45(3), 301 - 318.

Anderson, M. J. and Walsh, D.C.I. (2003). PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions: What null hypothesis are you testing? *Ecological Monographs*, 83(4), 557 - 574. Ecological Society of America.

Anderson, M. J. and ter Braak, C. J. F. (2003). Permutation tests for multi-factorial analysis of variance. *Journal of Statistical Computation and Simulation* 73, 85 - 113.

Anderson, M. J., Gorley, R. N., and Clarke, K. R. (2008). PERMANOVA + for PRIMER: *Guide to Software and Statistical Methods.* Plymouth, U.K.: PRIMER - E Ltd.

Dwass, M.(1957). Modified randomization tests for nonparametric hypotheses. Ann. Math. Statist. 28, 181-187.

Edgington, E. S. (1995). Randomization Tests, 3rd ed. Marcel Dekker, New York.

Everitt, B.S., Dunn, G. (1991), Applied Multivariate Data Analysis, Edward Arnold, London.

Gorley, R., Clarke,B. and Anderson, M.J. (PRIMER-E Ltd, Plymouth, UK)

Jari Oksanen, F. Guillaume Blanchet, Roeland Kindt, Pierre Legendre, Peter R. Minchin, R. B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens and Helene Wagner (2011). vegan: Community Ecology Package. R package version 2.0-1. http://CRAN.R-project.org/package=vegan

Kolb, G.S., Erkholm, J. Hambaeck, P.A. (2010). Effects of seabird nesting colonies on algae and aquatic invertebrates in coastal waters. *Marine Ecology Progress Series*, 417, 287-300.

Legendre P. and Anderson M. J. (1999). Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. *Ecol. Monogr.* 69, 1 - 24.

Legendre, P. and Legendre,L. (1998). Numerical Ecology. Elsevier, Amsterdam, The Netherlands.

Manly, B. F. J. (1997). Randomization, Bootstrap and Monte Carlo Methods in Biology, 2nd ed. Chapman and Hall, London.

Maxwell, S.E. and Delaney, H.D. (2004). Designing Experiments and Analizing Data, A Model Comparison Approach, 2nd ed. Lawerence Erlbaum Associates, London. McArdle, B. H. and Anderson, M. J. (2001). Fitting multivariate models to community data: a comment on distancebased redundancy analysis. *Ecology*, 82, 290 - 297.

Mielke, P.W.Jr. and Berry, K.J (2001). Permutation methods, a distance function approach, Srpinger Verlag, New York.

Montgomery, D.C. (2009). Design and analysis of experiments, 7th ed. John Wiley & Sons Inc., Asia.

R Development Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Reiss, P.T, Stevens, M.H.H., Shehzad, Z., Petkova, E. and Milham, M. (2001). On Dis-

tance - Based Permutation Tests for Between - Group Comparisons. *Biometrics*, 66, 636 - 643.
Searle, S.R. (1971), Linear Models, John Wiley & Sons, New York.
Searle, S.R. (1987), Linear Models for Unabalanced Data, John Wiley & Sons, New York.
Searle, S.R., Casella, G., McCulloch, C.E. (1992), Variance Components, John Wiley & Sons, New Jersey.

Part III
Appendix

Distribution	Unbalancedness	k	N	Distance	rejections.A	rejections.B	evalA1	evalA2
MVN	B (with no empty cells)	1.0000	60	Euclidean	0.0569	0.0649	closed	closed
MVN	B (with no empty cells)	3.0000	60	Euclidean	0.0589	0.0729	closed	liberal
MVN	B (with no empty cells)	10.0000	60	Euclidean	0.0689	0.0579	liberal	closed
MVN	B (with no empty cells)	20.0000	60	Euclidean	0.0649	0.0609	closed	closed
MVN	B (with 1 empty cell)	1.0000	60	Euclidean	0.0320	0.0609	conservative	closed
MVN	B (with 1 empty cell)	3.0000	60	Euclidean	0.0120	0.1179	conservative	liberal
MVN	B (with 1 empty cell)	10.0000	60	Euclidean	0.0130	0.1469	conservative	liberal
MVN	B (with 1 empty cell)	20.0000	60	Euclidean	0.0120	0.1459	conservative	liberal
MVN	B (with 2 empty cells)	1.0000	60	Euclidean	0.0569	0.0599	closed	closed
MVN	B (with 2 empty cells)	3.0000	60	Euclidean	0.0579	0.0659	closed	liberal
MVN	B (with 2 empty cells)	10.0000	60	Euclidean	0.0689	0.0559	liberal	closed
MVN	B (with 2 empty cells)	20.0000	60	Euclidean	0.0639	0.0579	closed	closed
MVN	U (with no empty cells)	1.0000	60	Euclidean	0.0519	0.0470	closed	closed
MVN	U (with no empty cells)	3.0000	60	Euclidean	0.1259	0.0519	liberal	closed
MVN	U (with no empty cells)	10.0000	60	Euclidean	0.2178	0.0609	liberal	closed
MVN	U (with no empty cells)	20.0000	60	Euclidean	0.2478	0.0639	liberal	closed
MVN	U (with no empty cells N2.>N1.)	1.0000	60	Euclidean	0.0440	0.0500	closed	closed
MVN	U (with no empty cells N2.>N1.)	3.0000	60	Euclidean	0.0010	0.0549	conservative	closed
MVN	U (with no empty cells N2.>N1.)	10.0000	60	Euclidean	0.0010	0.0539	conservative	closed
MVN	U (with no empty cells N2.>N1.)	20.0000	60	Euclidean	0.0010	0.0549	conservative	closed
MVN	U (with no empty cells N2.<N1.)	1.0000	60	Euclidean	0.0430	0.0549	closed	closed
MVN	U (with no empty cells N2.<N1.)	3.0000	60	Euclidean	0.3097	0.0689	liberal	liberal
MVN	U (with no empty cells N2.<N1.)	10.0000	60	Euclidean	0.5624	0.0939	liberal	liberal
MVN	U (with no empty cells N2.<N1.)	20.0000	60	Euclidean	0.6444	0.1129	liberal	liberal
MVN	U (with 1 empty cell)	1.0000	60	Euclidean	0.0420	0.0559	closed	closed
MVN	U (with 1 empty cell)	3.0000	60	Euclidean	0.0350	0.0899	conservative	liberal
MVN	U (with 1 empty cell)	10.0000	60	Euclidean	0.0350	0.1099	conservative	liberal
MVN	U (with 1 empty cell)	20.0000	60	Euclidean	0.0390	0.1189	closed	liberal
MVN	U (with 2 empty cells)	1.0000	60	Euclidean	0.0410	0.0529	closed	closed
MVN	U (with 2 empty cells)	3.0000	60	Euclidean	0.0090	0.0849	conservative	liberal
MVN	U (with 2 empty cells)	10.0000	60	Euclidean	0.0010	0.1109	conservative	liberal
MVN	U (with 2 empty cells)	20.0000	60	Euclidean	0.0010	0.1169	conservative	liberal
MVN	U (with 3 empty cells)	1.0000	60	Euclidean	0.0440	0.0490	closed	closed
MVN	U (with 3 empty cells)	3.0000	60	Euclidean	0.0130	0.1249	conservative	liberal
MVN	U (with 3 empty cells)	10.0000	60	Euclidean	0.0090	0.1768	conservative	liberal
MVN	U (with 3 empty cells)	20.0000	60	Euclidean	0.0080	0.1888	conservative	liberal
MVN	B (with no empty cells)	1.0000	120	Euclidean	0.0380	0.0440	closed	closed
MVN	B (with no empty cells)	3.0000	120	Euclidean	0.0460	0.0480	closed	closed
MVN	B (with no empty cells)	10.0000	120	Euclidean	0.0549	0.0559	closed	closed
MVN	B (with no empty cells)	20.0000	120	Euclidean	0.0569	0.0559	closed	closed
MVN	B (with 1 empty cell)	1.0000	120	Euclidean	0.0420	0.0460	closed	closed
MVN	B (with 1 empty cell)	3.0000	120	Euclidean	0.0220	0.0949	conservative	liberal
MVN	B (with 1 empty cell)	10.0000	120	Euclidean	0.0140	0.1209	conservative	liberal
MVN	B (with 1 empty cell)	20.0000	120	Euclidean	0.0140	0.1319	conservative	liberal
MVN	B (with 2 empty cells)	1.0000	120	Euclidean	0.0370	0.0490	closed	closed
MVN	B (with 2 empty cells)	3.0000	120	Euclidean	0.0470	0.0549	closed	closed
MVN	B (with 2 empty cells)	10.0000	120	Euclidean	0.0569	0.0559	closed	closed
MVN	B (with 2 empty cells)	20.0000	120	Euclidean	0.0589	0.0569	closed	closed
MVN	U (with no empty cells)	1.0000	120	Euclidean	0.0310	0.0420	conservative	closed
MVN	U (with no empty cells)	3.0000	120	Euclidean	0.0969	0.0589	liberal	closed
MVN	U (with no empty cells)	10.0000	120	Euclidean	0.1608	0.0649	liberal	closed
MVN	U (with no empty cells)	20.0000	120	Euclidean	0.1818	0.0659	liberal	liberal
MVN	U (with no empty cells N2.>N1.)	1.0000	120	Euclidean	0.0519	0.0410	closed	closed
MVN	U (with no empty cells N2.>N1.)	3.0000	120	Euclidean	0.0020	0.0420	conservative	closed
MVN	U (with no empty cells N2.>N1.)	10.0000	120	Euclidean	0.0010	0.0460	conservative	closed
MVN	U (with no empty cells N2.>N1.)	20.0000	120	Euclidean	0.0010	0.0460	conservative	closed
MVN	U (with no empty cells N2.<N1.)	1.0000	120	Euclidean	0.0559	0.0410	closed	closed
MVN	U (with no empty cells N2.<N1.)	3.0000	120	Euclidean	0.4775	0.0470	liberal	closed
MVN	U (with no empty cells N2.<N1.)	10.0000	120	Euclidean	0.8092	0.0639	liberal	closed
MVN	U (with no empty cells N2.<N1.)	20.0000	120	Euclidean	0.8691	0.0739	liberal	liberal
MVN	U (with 1 empty cell)	1.0000	120	Euclidean	0.0529	0.0390	closed	closed
MVN	U (with 1 empty cell)	3.0000	120	Euclidean	0.0350	0.0799	conservative	liberal
MVN	U (with 1 empty cell)	10.0000	120	Euclidean	0.0280	0.1139	conservative	liberal
MVN	U (with 1 empty cell)	20.0000	120	Euclidean	0.0230	0.1259	conservative	liberal
MVN	U (with 2 empty cells)	1.0000	120	Euclidean	0.0509	0.0559	closed	closed
MVN	U (with 2 empty cells)	3.0000	120	Euclidean	0.0100	0.0909	conservative	liberal
MVN	U (with 2 empty cells)	10.0000	120	Euclidean	0.0020	0.1189	conservative	liberal
MVN	U (with 2 empty cells)	20.0000	120	Euclidean	0.0020	0.1279	conservative	liberal
MVN	U (with 3 empty cells)	1.0000	120	Euclidean	0.0539	0.0410	closed	closed
MVN	U (with 3 empty cells)	3.0000	120	Euclidean	0.0220	0.1289	conservative	liberal
MVN	U (with 3 empty cells)	10.0000	120	Euclidean	0.0060	0.1648	conservative	liberal
MVN	U (with 3 empty cells)	20.0000	120	Euclidean	0.0060	0.1718	conservative	liberal

Table 25: Results (normal variates under Euclidean measure when $p = 4$). The delimited results correspond to scenarios that were similar to those ecnountered in the real case.

Distribution	Unbalancedness	k	N	Distance	rejections.A	rejections.B	evalA1	evalA2
Nbin	B (with no empty cells)	1.0000	60	Euclidean	0.0529	0.0599	closed	closed
Nbin	B (with no empty cells)	0.0500	60	Euclidean	0.0519	0.0539	closed	closed
Nbin	B (with no empty cells)	0.0200	60	Euclidean	0.0529	0.0559	closed	closed
Nbin	B (with no empty cells)	0.0100	60	Euclidean	0.0490	0.0480	closed	closed
Nbin	B (with 1 empty cell)	1.0000	60	Euclidean	0.0559	0.0450	closed	closed
Nbin	B (with 1 empty cell)	0.0500	60	Euclidean	0.0420	0.0649	closed	closed
Nbin	B (with 1 empty cell)	0.0200	60	Euclidean	0.0320	0.0769	conservative	liberal
Nbin	B (with 1 empty cell)	0.0100	60	Euclidean	0.0330	0.0949	conservative	liberal
Nbin	B (with 2 empty cells)	1.0000	60	Euclidean	0.0509	0.0509	closed	closed
Nbin	B (with 2 empty cells)	0.0500	60	Euclidean	0.0509	0.0500	closed	closed
Nbin	B (with 2 empty cells)	0.0200	60	Euclidean	0.0509	0.0450	closed	closed
Nbin	B (with 2 empty cells)	0.0100	60	Euclidean	0.0500	0.0500	closed	closed
Nbin	U (with no empty cells)	1.0000	60	Euclidean	0.0519	0.0609	closed	closed
Nbin	U (with no empty cells)	0.0500	60	Euclidean	0.0639	0.0509	closed	closed
Nbin	U (with no empty cells)	0.0200	60	Euclidean	0.0699	0.0470	liberal	closed
Nbin	U (with no empty cells)	0.0100	60	Euclidean	0.0969	0.0659	liberal	liberal
Nbin	U (with no empty cells N2.>N1.)	1.0000	60	Euclidean	0.0500	0.0440	closed	closed
Nbin	U (with no empty cells N2.>N1.)	0.0500	60	Euclidean	0.0210	0.0490	conservative	closed
Nbin	U (with no empty cells N2.>N1.)	0.0200	60	Euclidean	0.0120	0.0649	conservative	closed
Nbin	U (with no empty cells N2.>N1.)	0.0100	60	Euclidean	0.0070	0.0629	conservative	closed
Nbin	U (with no empty cells N2.<N1.)	1.0000	60	Euclidean	0.0549	0.0420	closed	closed
Nbin	U (with no empty cells N2.<N1.)	0.0500	60	Euclidean	0.0719	0.0470	liberal	closed
Nbin	U (with no empty cells N2.<N1.)	0.0200	60	Euclidean	0.1319	0.0569	liberal	closed
Nbin	U (with no empty cells N2.<N1.)	0.0100	60	Euclidean	0.2038	0.0679	liberal	liberal
Nbin	U (with 1 empty cell)	1.0000	60	Euclidean	0.0490	0.0410	closed	closed
Nbin	U (with 1 empty cell)	0.0500	60	Euclidean	0.0490	0.0589	closed	closed
Nbin	U (with 1 empty cell)	0.0200	60	Euclidean	0.0460	0.0589	closed	closed
Nbin	U (with 1 empty cell)	0.0100	60	Euclidean	0.0420	0.0859	closed	liberal
Nbin	U (with 2 empty cells)	1.0000	60	Euclidean	0.0460	0.0500	closed	closed
Nbin	U (with 2 empty cells)	0.0500	60	Euclidean	0.0370	0.0400	closed	closed
Nbin	U (with 2 empty cells)	0.0200	60	Euclidean	0.0150	0.0629	conservative	closed
Nbin	U (with 2 empty cells)	0.0100	60	Euclidean	0.0070	0.1049	conservative	liberal
Nbin	U (with 3 empty cells)	1.0000	60	Euclidean	0.0559	0.0460	closed	closed
Nbin	U (with 3 empty cells)	0.0500	60	Euclidean	0.0380	0.0729	closed	liberal
Nbin	U (with 3 empty cells)	0.0200	60	Euclidean	0.0310	0.0739	conservative	liberal
Nbin	U (with 3 empty cells)	0.0100	60	Euclidean	0.0240	0.0929	conservative	liberal
Nbin	B (with no empty cells)	1.0000	120	Euclidean	0.0440	0.0490	closed	closed
Nbin	B (with no empty cells)	0.0500	120	Euclidean	0.0490	0.0589	closed	closed
Nbin	B (with no empty cells)	0.0200	120	Euclidean	0.0519	0.0559	closed	closed
Nbin	B (with no empty cells)	0.0100	120	Euclidean	0.0490	0.0500	closed	closed
Nbin	B (with 1 empty cell)	1.0000	120	Euclidean	0.0539	0.0519	closed	closed
Nbin	B (with 1 empty cell)	0.0500	120	Euclidean	0.0410	0.0709	closed	liberal
Nbin	B (with 1 empty cell)	0.0200	120	Euclidean	0.0360	0.0440	conservative	closed
Nbin	B (with 1 empty cell)	0.0100	120	Euclidean	0.0300	0.0699	conservative	liberal
Nbin	B (with 2 empty cells)	1.0000	120	Euclidean	0.0430	0.0490	closed	closed
Nbin	B (with 2 empty cells)	0.0500	120	Euclidean	0.0470	0.0549	closed	closed
Nbin	B (with 2 empty cells)	0.0200	120	Euclidean	0.0509	0.0500	closed	closed
Nbin	B (with 2 empty cells)	0.0100	120	Euclidean	0.0480	0.0430	closed	closed
Nbin	U (with no empty cells)	1.0000	120	Euclidean	0.0509	0.0480	closed	closed
Nbin	U (with no empty cells)	0.0500	120	Euclidean	0.0470	0.0659	closed	liberal
Nbin	U (with no empty cells)	0.0200	120	Euclidean	0.0549	0.0480	closed	closed
Nbin	U (with no empty cells)	0.0100	120	Euclidean	0.0929	0.0509	liberal	closed
Nbin	U (with no empty cells N2.>N1.)	1.0000	120	Euclidean	0.0519	0.0480	closed	closed
Nbin	U (with no empty cells N2.>N1.)	0.0500	120	Euclidean	0.0310	0.0440	conservative	closed
Nbin	U (with no empty cells N2.>N1.)	0.0200	120	Euclidean	0.0150	0.0619	conservative	closed
Nbin	U (with no empty cells N2.>N1.)	0.0100	120	Euclidean	0.0010	0.0430	conservative	closed
Nbin	U (with no empty cells N2.<N1.)	1.0000	120	Euclidean	0.0480	0.0480	closed	closed
Nbin	U (with no empty cells N2.<N1.)	0.0500	120	Euclidean	0.0869	0.0500	liberal	closed
Nbin	U (with no empty cells N2.<N1.)	0.0200	120	Euclidean	0.1678	0.0410	liberal	closed
Nbin	U (with no empty cells N2.<N1.)	0.0100	120	Euclidean	0.2967	0.0490	liberal	closed
Nbin	U (with 1 empty cell)	1.0000	120	Euclidean	0.0559	0.0559	closed	closed
Nbin	U (with 1 empty cell)	0.0500	120	Euclidean	0.0350	0.0519	conservative	closed
Nbin	U (with 1 empty cell)	0.0200	120	Euclidean	0.0440	0.0749	closed	liberal
Nbin	U (with 1 empty cell)	0.0100	120	Euclidean	0.0470	0.0679	closed	liberal
Nbin	U (with 2 empty cells)	1.0000	120	Euclidean	0.0569	0.0559	closed	closed
Nbin	U (with 2 empty cells)	0.0500	120	Euclidean	0.0480	0.0679	closed	liberal
Nbin	U (with 2 empty cells)	0.0200	120	Euclidean	0.0120	0.0609	conservative	closed
Nbin	U (with 2 empty cells)	0.0100	120	Euclidean	0.0070	0.0939	conservative	liberal
Nbin	U (with 3 empty cells)	1.0000	120	Euclidean	0.0529	0.0539	closed	closed
Nbin	U (with 3 empty cells)	0.0500	120	Euclidean	0.0410	0.0729	closed	liberal
Nbin	U (with 3 empty cells)	0.0200	120	Euclidean	0.0260	0.0799	conservative	liberal
Nbin	U (with 3 empty cells)	0.0100	120	Euclidean	0.0250	0.0969	conservative	liberal

Table 26: Results (negative binomial variates under Euclidean measure when $p = 4$). The delimited results correspond to scenarios that were similar to those ecnountered in the real case.

Distribution	Unbalancedness	k	N	Distance	rejections.A	rejections.B	evalA1	evalA2
Nbin	B (with no empty cells)	1.0000	60	Bray Curtis	0.0480	0.0529	closed	closed
Nbin	B (with no empty cells)	0.0500	60	Bray Curtis	0.0390	0.0509	closed	closed
Nbin	B (with no empty cells)	0.0200	60	Bray Curtis	0.0529	0.0509	closed	closed
Nbin	B (with no empty cells)	0.0100	60	Bray Curtis	0.0480	0.0440	closed	closed
Nbin	B (with 1 empty cell)	1.0000	60	Bray Curtis	0.0509	0.0519	closed	closed
Nbin	B (with 1 empty cell)	0.0500	60	Bray Curtis	0.0440	0.0579	closed	closed
Nbin	B (with 1 empty cell)	0.0200	60	Bray Curtis	0.0559	0.0579	closed	closed
Nbin	B (with 1 empty cell)	0.0100	60	Bray Curtis	0.0460	0.0609	closed	closed
Nbin	B (with 2 empty cells)	1.0000	60	Bray Curtis	0.0480	0.0519	closed	closed
Nbin	B (with 2 empty cells)	0.0500	60	Bray Curtis	0.0410	0.0539	closed	closed
Nbin	B (with 2 empty cells)	0.0200	60	Bray Curtis	0.0529	0.0490	closed	closed
Nbin	B (with 2 empty cells)	0.0100	60	Bray Curtis	0.0490	0.0509	closed	closed
Nbin	U (with no empty cells)	1.0000	60	Bray Curtis	0.0509	0.0559	closed	closed
Nbin	U (with no empty cells)	0.0500	60	Bray Curtis	0.0509	0.0480	closed	closed
Nbin	U (with no empty cells)	0.0200	60	Bray Curtis	0.0549	0.0490	closed	closed
Nbin	U (with no empty cells)	0.0100	60	Bray Curtis	0.0639	0.0460	closed	closed
Nbin	U (with no empty cells N2.>N1.)	1.0000	60	Bray Curtis	0.0470	0.0490	closed	closed
Nbin	U (with no empty cells N2.>N1.)	0.0500	60	Bray Curtis	0.0390	0.0539	closed	closed
Nbin	U (with no empty cells N2.>N1.)	0.0200	60	Bray Curtis	0.0410	0.0500	closed	closed
Nbin	U (with no empty cells N2.>N1.)	0.0100	60	Bray Curtis	0.0370	0.0470	closed	closed
Nbin	U (with no empty cells N2.<N1.)	1.0000	60	Bray Curtis	0.0599	0.0440	closed	closed
Nbin	U (with no empty cells N2.<N1.)	0.0500	60	Bray Curtis	0.0450	0.0450	closed	closed
Nbin	U (with no empty cells N2.<N1.)	0.0200	60	Bray Curtis	0.0569	0.0490	closed	closed
Nbin	U (with no empty cells N2.<N1.)	0.0100	60	Bray Curtis	0.0639	0.0529	closed	closed
Nbin	U (with 1 empty cell)	1.0000	60	Bray Curtis	0.0480	0.0420	closed	closed
Nbin	U (with 1 empty cell)	0.0500	60	Bray Curtis	0.0509	0.0400	closed	closed
Nbin	U (with 1 empty cell)	0.0200	60	Bray Curtis	0.0440	0.0450	closed	closed
Nbin	U (with 1 empty cell)	0.0100	60	Bray Curtis	0.0500	0.0579	closed	closed
Nbin	U (with 2 empty cells)	1.0000	60	Bray Curtis	0.0509	0.0500	closed	closed
Nbin	U (with 2 empty cells)	0.0500	60	Bray Curtis	0.0699	0.0559	liberal	closed
Nbin	U (with 2 empty cells)	0.0200	60	Bray Curtis	0.0470	0.0549	closed	closed
Nbin	U (with 2 empty cells)	0.0100	60	Bray Curtis	0.0370	0.0509	closed	closed
Nbin	U (with 3 empty cells)	1.0000	60	Bray Curtis	0.0559	0.0500	closed	closed
Nbin	U (with 3 empty cells)	0.0500	60	Bray Curtis	0.0470	0.0519	closed	closed
Nbin	U (with 3 empty cells)	0.0200	60	Bray Curtis	0.0460	0.0430	closed	closed
Nbin	U (with 3 empty cells)	0.0100	60	Bray Curtis	0.0480	0.0529	closed	closed
Nbin	B (with no empty cells)	1.0000	120	Bray Curtis	0.0500	0.0450	closed	closed
Nbin	B (with no empty cells)	0.0500	120	Bray Curtis	0.0440	0.0420	closed	closed
Nbin	B (with no empty cells)	0.0200	120	Bray Curtis	0.0589	0.0529	closed	closed
Nbin	B (with no empty cells)	0.0100	120	Bray Curtis	0.0519	0.0639	closed	closed
Nbin	B (with 1 empty cell)	1.0000	120	Bray Curtis	0.0400	0.0430	closed	closed
Nbin	B (with 1 empty cell)	0.0500	120	Bray Curtis	0.0559	0.0559	closed	closed
Nbin	B (with 1 empty cell)	0.0200	120	Bray Curtis	0.0589	0.0490	closed	closed
Nbin	B (with 1 empty cell)	0.0100	120	Bray Curtis	0.0350	0.0659	conservative	liberal
Nbin	B (with 2 empty cells)	1.0000	120	Bray Curtis	0.0500	0.0410	closed	closed
Nbin	B (with 2 empty cells)	0.0500	120	Bray Curtis	0.0440	0.0400	closed	closed
Nbin	B (with 2 empty cells)	0.0200	120	Bray Curtis	0.0589	0.0579	closed	closed
Nbin	B (with 2 empty cells)	0.0100	120	Bray Curtis	0.0539	0.0549	closed	closed
Nbin	U (with no empty cells)	1.0000	120	Bray Curtis	0.0539	0.0490	closed	closed
Nbin	U (with no empty cells)	0.0500	120	Bray Curtis	0.0519	0.0579	closed	closed
Nbin	U (with no empty cells)	0.0200	120	Bray Curtis	0.0470	0.0430	closed	closed
Nbin	U (with no empty cells)	0.0100	120	Bray Curtis	0.0549	0.0569	closed	closed
Nbin	U (with no empty cells N2.>N1.)	1.0000	120	Bray Curtis	0.0539	0.0420	closed	closed
Nbin	U (with no empty cells N2.>N1.)	0.0500	120	Bray Curtis	0.0549	0.0310	closed	conservative
Nbin	U (with no empty cells N2.>N1.)	0.0200	120	Bray Curtis	0.0410	0.0509	closed	closed
Nbin	U (with no empty cells N2.>N1.)	0.0100	120	Bray Curtis	0.0300	0.0380	conservative	closed
Nbin	U (with no empty cells N2.<N1.)	1.0000	120	Bray Curtis	0.0420	0.0370	closed	closed
Nbin	U (with no empty cells N2.<N1.)	0.0500	120	Bray Curtis	0.0579	0.0380	closed	closed
Nbin	U (with no empty cells N2.<N1.)	0.0200	120	Bray Curtis	0.0519	0.0390	closed	closed
Nbin	U (with no empty cells N2.<N1.)	0.0100	120	Bray Curtis	0.0659	0.0390	liberal	closed
Nbin	U (with 1 empty cell)	1.0000	120	Bray Curtis	0.0450	0.0629	closed	closed
Nbin	U (with 1 empty cell)	0.0500	120	Bray Curtis	0.0480	0.0410	closed	closed
Nbin	U (with 1 empty cell)	0.0200	120	Bray Curtis	0.0470	0.0579	closed	closed
Nbin	U (with 1 empty cell)	0.0100	120	Bray Curtis	0.0549	0.0440	closed	closed
Nbin	U (with 2 empty cells)	1.0000	120	Bray Curtis	0.0589	0.0539	closed	closed
Nbin	U (with 2 empty cells)	0.0500	120	Bray Curtis	0.0490	0.0609	closed	closed
Nbin	U (with 2 empty cells)	0.0200	120	Bray Curtis	0.0509	0.0559	closed	closed
Nbin	U (with 2 empty cells)	0.0100	120	Bray Curtis	0.0509	0.0579	closed	closed
Nbin	U (with 3 empty cells)	1.0000	120	Bray Curtis	0.0529	0.0599	closed	closed
Nbin	U (with 3 empty cells)	0.0500	120	Bray Curtis	0.0579	0.0579	closed	closed
Nbin	U (with 3 empty cells)	0.0200	120	Bray Curtis	0.0430	0.0460	closed	closed
Nbin	U (with 3 empty cells)	0.0100	120	Bray Curtis	0.0490	0.0549	closed	closed

Table 27: Results (negative binomial variates under Bray Curtis measure when $p = 4$).

Distribution	Unbalancedness	k	N	Distance	rejections.A	rejections.B	Factor A	Factor B
MVN	B (with no empty cells)	1.0000	60	Euclidean	0.0500	0.0490	closed	closed
MVN	B (with no empty cells)	3.0000	60	Euclidean	0.0480	0.0549	closed	closed
MVN	B (with no empty cells)	10.0000	60	Euclidean	0.0549	0.0519	closed	closed
MVN	B (with no empty cells)	20.0000	60	Euclidean	0.0539	0.0579	closed	closed
MVN	B (with 1 empty cell)	1.0000	60	Euclidean	0.0490	0.0440	closed	closed
MVN	B (with 1 empty cell)	3.0000	60	Euclidean	0.0220	0.0829	conservative	liberal
MVN	B (with 1 empty cell)	10.0000	60	Euclidean	0.0100	0.1099	conservative	liberal
MVN	B (with 1 empty cell)	20.0000	60	Euclidean	0.0070	0.1119	conservative	liberal
MVN	B (with 2 empty cells)	1.0000	60	Euclidean	0.0509	0.0440	closed	closed
MVN	B (with 2 empty cells)	3.0000	60	Euclidean	0.0480	0.0529	closed	closed
MVN	B (with 2 empty cells)	10.0000	60	Euclidean	0.0539	0.0480	closed	closed
MVN	B (with 2 empty cells)	20.0000	60	Euclidean	0.0549	0.0509	closed	closed
MVN	U (with no empty cells)	1.0000	60	Euclidean	0.0490	0.0549	closed	closed
MVN	U (with no empty cells)	3.0000	60	Euclidean	0.1379	0.0699	liberal	liberal
MVN	U (with no empty cells)	10.0000	60	Euclidean	0.2278	0.0809	liberal	liberal
MVN	U (with no empty cells)	20.0000	60	Euclidean	0.2517	0.0769	liberal	liberal
MVN	U (with no empty cells N2.>N1.)	1.0000	60	Euclidean	0.0529	0.0420	closed	closed
MVN	U (with no empty cells N2.>N1.)	3.0000	60	Euclidean	0.0010	0.0490	conservative	closed
MVN	U (with no empty cells N2.>N1.)	10.0000	60	Euclidean	0.0010	0.0440	conservative	closed
MVN	U (with no empty cells N2.>N1.)	20.0000	60	Euclidean	0.0010	0.0450	conservative	closed
MVN	U (with no empty cells N2.<N1.)	1.0000	60	Euclidean	0.0500	0.0380	closed	closed
MVN	U (with no empty cells N2.<N1.)	3.0000	60	Euclidean	0.3287	0.0589	liberal	closed
MVN	U (with no empty cells N2.<N1.)	10.0000	60	Euclidean	0.5964	0.1089	liberal	liberal
MVN	U (with no empty cells N2.<N1.)	20.0000	60	Euclidean	0.6633	0.1169	liberal	liberal
MVN	U (with 1 empty cell)	1.0000	60	Euclidean	0.0450	0.0629	closed	closed
MVN	U (with 1 empty cell)	3.0000	60	Euclidean	0.0350	0.0919	conservative	liberal
MVN	U (with 1 empty cell)	10.0000	60	Euclidean	0.0300	0.1039	conservative	liberal
MVN	U (with 1 empty cell)	20.0000	60	Euclidean	0.0280	0.1109	conservative	liberal
MVN	U (with 2 empty cells)	1.0000	60	Euclidean	0.0599	0.0480	closed	closed
MVN	U (with 2 empty cells)	3.0000	60	Euclidean	0.0060	0.0829	conservative	liberal
MVN	U (with 2 empty cells)	10.0000	60	Euclidean	0.0020	0.1029	conservative	liberal
MVN	U (with 2 empty cells)	20.0000	60	Euclidean	0.0020	0.1089	conservative	liberal
MVN	U (with 3 empty cells)	1.0000	60	Euclidean	0.0420	0.0460	closed	closed
MVN	U (with 3 empty cells)	3.0000	60	Euclidean	0.0080	0.1269	conservative	liberal
MVN	U (with 3 empty cells)	10.0000	60	Euclidean	0.0050	0.1648	conservative	liberal
MVN	U (with 3 empty cells)	20.0000	60	Euclidean	0.0040	0.1748	conservative	liberal
MVN	B (with no empty cells)	1.0000	120	Euclidean	0.0420	0.0360	closed	conservative
MVN	B (with no empty cells)	3.0000	120	Euclidean	0.0470	0.0380	closed	closed
MVN	B (with no empty cells)	10.0000	120	Euclidean	0.0539	0.0410	closed	closed
MVN	B (with no empty cells)	20.0000	120	Euclidean	0.0549	0.0410	closed	closed
MVN	B (with 1 empty cell)	1.0000	120	Euclidean	0.0539	0.0440	closed	closed
MVN	B (with 1 empty cell)	3.0000	120	Euclidean	0.0320	0.0869	conservative	liberal
MVN	B (with 1 empty cell)	10.0000	120	Euclidean	0.0170	0.1039	conservative	liberal
MVN	B (with 1 empty cell)	20.0000	120	Euclidean	0.0130	0.1099	conservative	liberal
MVN	B (with 2 empty cells)	1.0000	120	Euclidean	0.0420	0.0330	closed	conservative
MVN	B (with 2 empty cells)	3.0000	120	Euclidean	0.0470	0.0410	closed	closed
MVN	B (with 2 empty cells)	10.0000	120	Euclidean	0.0539	0.0440	closed	closed
MVN	B (with 2 empty cells)	20.0000	120	Euclidean	0.0539	0.0480	closed	closed
MVN	U (with no empty cells)	1.0000	120	Euclidean	0.0569	0.0579	closed	closed
MVN	U (with no empty cells)	3.0000	120	Euclidean	0.1109	0.0709	liberal	liberal
MVN	U (with no empty cells)	10.0000	120	Euclidean	0.1698	0.0779	liberal	liberal
MVN	U (with no empty cells)	20.0000	120	Euclidean	0.1848	0.0829	liberal	liberal
MVN	U (with no empty cells N2.>N1.)	1.0000	120	Euclidean	0.0420	0.0430	closed	closed
MVN	U (with no empty cells N2.>N1.)	3.0000	120	Euclidean	0.0030	0.0430	conservative	closed
MVN	U (with no empty cells N2.>N1.)	10.0000	120	Euclidean	0.0010	0.0410	conservative	closed
MVN	U (with no empty cells N2.>N1.)	20.0000	120	Euclidean	0.0010	0.0450	conservative	closed
MVN	U (with no empty cells N2.<N1.)	1.0000	120	Euclidean	0.0480	0.0559	closed	closed
MVN	U (with no empty cells N2.<N1.)	3.0000	120	Euclidean	0.4765	0.0539	liberal	closed
MVN	U (with no empty cells N2.<N1.)	10.0000	120	Euclidean	0.8122	0.0679	liberal	liberal
MVN	U (with no empty cells N2.<N1.)	20.0000	120	Euclidean	0.8741	0.0869	liberal	liberal
MVN	U (with 1 empty cell)	1.0000	120	Euclidean	0.0500	0.0589	closed	closed
MVN	U (with 1 empty cell)	3.0000	120	Euclidean	0.0280	0.0959	conservative	liberal
MVN	U (with 1 empty cell)	10.0000	120	Euclidean	0.0160	0.1189	conservative	liberal
MVN	U (with 1 empty cell)	20.0000	120	Euclidean	0.0170	0.1229	conservative	liberal
MVN	U (with 2 empty cells)	1.0000	120	Euclidean	0.0500	0.0470	closed	closed
MVN	U (with 2 empty cells)	3.0000	120	Euclidean	0.0040	0.1079	conservative	liberal
MVN	U (with 2 empty cells)	10.0000	120	Euclidean	0.0020	0.1279	conservative	liberal
MVN	U (with 2 empty cells)	20.0000	120	Euclidean	0.0010	0.1299	conservative	liberal
MVN	U (with 3 empty cells)	1.0000	120	Euclidean	0.0490	0.0420	closed	closed
MVN	U (with 3 empty cells)	3.0000	120	Euclidean	0.0140	0.1089	conservative	liberal
MVN	U (with 3 empty cells)	10.0000	120	Euclidean	0.0030	0.1499	conservative	liberal
MVN	U (with 3 empty cells)	20.0000	120	Euclidean	0.0030	0.1548	conservative	liberal

Table 28: Results (normal variates under Euclidean measure when $p = 150$).

Distribution	Unbalancedness	k	N	Distance	rejections.A	rejections.B	Factor A	Factor B
Nbin	B (with no empty cells)	1.0000	60	Euclidean	0.0609	0.0420	closed	closed
Nbin	B (with no empty cells)	0.0500	60	Euclidean	0.0410	0.0460	closed	closed
Nbin	B (with no empty cells)	0.0200	60	Euclidean	0.0619	0.0549	closed	closed
Nbin	B (with no empty cells)	0.0100	60	Euclidean	0.0450	0.0569	closed	closed
Nbin	B (with 1 empty cell)	1.0000	60	Euclidean	0.0669	0.0390	liberal	closed
Nbin	B (with 1 empty cell)	0.0500	60	Euclidean	0.0410	0.0559	closed	closed
Nbin	B (with 1 empty cell)	0.0200	60	Euclidean	0.0310	0.0599	conservative	closed
Nbin	B (with 1 empty cell)	0.0100	60	Euclidean	0.0230	0.0709	conservative	liberal
Nbin	B (with 2 empty cells)	1.0000	60	Euclidean	0.0619	0.0509	closed	closed
Nbin	B (with 2 empty cells)	0.0500	60	Euclidean	0.0440	0.0380	closed	closed
Nbin	B (with 2 empty cells)	0.0200	60	Euclidean	0.0619	0.0539	closed	closed
Nbin	B (with 2 empty cells)	0.0100	60	Euclidean	0.0450	0.0500	closed	closed
Nbin	U (with no empty cells)	1.0000	60	Euclidean	0.0559	0.0460	closed	closed
Nbin	U (with no empty cells)	0.0500	60	Euclidean	0.0709	0.0549	liberal	closed
Nbin	U (with no empty cells)	0.0200	60	Euclidean	0.0919	0.0470	liberal	closed
Nbin	U (with no empty cells)	0.0100	60	Euclidean	0.1049	0.0619	liberal	closed
Nbin	U (with no empty cells N2.>N1.)	1.0000	60	Euclidean	0.0549	0.0410	closed	closed
Nbin	U (with no empty cells N2.>N1.)	0.0500	60	Euclidean	0.0250	0.0440	conservative	closed
Nbin	U (with no empty cells N2.>N1.)	0.0200	60	Euclidean	0.0160	0.0529	conservative	closed
Nbin	U (with no empty cells N2.>N1.)	0.0100	60	Euclidean	0.0050	0.0509	conservative	closed
Nbin	U (with no empty cells N2.<N1.)	1.0000	60	Euclidean	0.0539	0.0549	closed	closed
Nbin	U (with no empty cells N2.<N1.)	0.0500	60	Euclidean	0.0949	0.0589	liberal	closed
Nbin	U (with no empty cells N2.<N1.)	0.0200	60	Euclidean	0.1449	0.0519	liberal	closed
Nbin	U (with no empty cells N2.<N1.)	0.0100	60	Euclidean	0.2218	0.0659	liberal	liberal
Nbin	U (with 1 empty cell)	1.0000	60	Euclidean	0.0549	0.0589	closed	closed
Nbin	U (with 1 empty cell)	0.0500	60	Euclidean	0.0579	0.0509	closed	closed
Nbin	U (with 1 empty cell)	0.0200	60	Euclidean	0.0360	0.0669	conservative	liberal
Nbin	U (with 1 empty cell)	0.0100	60	Euclidean	0.0490	0.0749	closed	liberal
Nbin	U (with 2 empty cells)	1.0000	60	Euclidean	0.0390	0.0509	closed	closed
Nbin	U (with 2 empty cells)	0.0500	60	Euclidean	0.0330	0.0589	conservative	closed
Nbin	U (with 2 empty cells)	0.0200	60	Euclidean	0.0160	0.0609	conservative	closed
Nbin	U (with 2 empty cells)	0.0100	60	Euclidean	0.0080	0.0899	conservative	liberal
Nbin	U (with 3 empty cells)	1.0000	60	Euclidean	0.0549	0.0579	closed	closed
Nbin	U (with 3 empty cells)	0.0500	60	Euclidean	0.0350	0.0609	conservative	closed
Nbin	U (with 3 empty cells)	0.0200	60	Euclidean	0.0260	0.0919	conservative	liberal
Nbin	U (with 3 empty cells)	0.0100	60	Euclidean	0.0320	0.1059	conservative	liberal
Nbin	B (with no empty cells)	1.0000	120	Euclidean	0.0559	0.0430	closed	closed
Nbin	B (with no empty cells)	0.0500	120	Euclidean	0.0529	0.0430	closed	closed
Nbin	B (with no empty cells)	0.0200	120	Euclidean	0.0470	0.0460	closed	closed
Nbin	B (with no empty cells)	0.0100	120	Euclidean	0.0470	0.0340	closed	conservative
Nbin	B (with 1 empty cell)	1.0000	120	Euclidean	0.0519	0.0599	closed	closed
Nbin	B (with 1 empty cell)	0.0500	120	Euclidean	0.0450	0.0559	closed	closed
Nbin	B (with 1 empty cell)	0.0200	120	Euclidean	0.0330	0.0529	conservative	closed
Nbin	B (with 1 empty cell)	0.0100	120	Euclidean	0.0260	0.0889	conservative	liberal
Nbin	B (with 2 empty cells)	1.0000	120	Euclidean	0.0569	0.0559	closed	closed
Nbin	B (with 2 empty cells)	0.0500	120	Euclidean	0.0529	0.0490	closed	closed
Nbin	B (with 2 empty cells)	0.0200	120	Euclidean	0.0480	0.0599	closed	closed
Nbin	B (with 2 empty cells)	0.0100	120	Euclidean	0.0470	0.0430	closed	closed
Nbin	U (with no empty cells)	1.0000	120	Euclidean	0.0639	0.0410	closed	closed
Nbin	U (with no empty cells)	0.0500	120	Euclidean	0.0559	0.0559	closed	closed
Nbin	U (with no empty cells)	0.0200	120	Euclidean	0.0579	0.0539	closed	closed
Nbin	U (with no empty cells)	0.0100	120	Euclidean	0.0839	0.0589	liberal	closed
Nbin	U (with no empty cells N2.>N1.)	1.0000	120	Euclidean	0.0490	0.0509	closed	closed
Nbin	U (with no empty cells N2.>N1.)	0.0500	120	Euclidean	0.0300	0.0569	conservative	closed
Nbin	U (with no empty cells N2.>N1.)	0.0200	120	Euclidean	0.0160	0.0450	conservative	closed
Nbin	U (with no empty cells N2.>N1.)	0.0100	120	Euclidean	0.0030	0.0490	conservative	closed
Nbin	U (with no empty cells N2.<N1.)	1.0000	120	Euclidean	0.0440	0.0470	closed	closed
Nbin	U (with no empty cells N2.<N1.)	0.0500	120	Euclidean	0.0869	0.0609	liberal	closed
Nbin	U (with no empty cells N2.<N1.)	0.0200	120	Euclidean	0.1698	0.0519	liberal	closed
Nbin	U (with no empty cells N2.<N1.)	0.0100	120	Euclidean	0.3267	0.0509	liberal	closed
Nbin	U (with 1 empty cell)	1.0000	120	Euclidean	0.0619	0.0509	closed	closed
Nbin	U (with 1 empty cell)	0.0500	120	Euclidean	0.0460	0.0679	closed	liberal
Nbin	U (with 1 empty cell)	0.0200	120	Euclidean	0.0320	0.0519	conservative	closed
Nbin	U (with 1 empty cell)	0.0100	120	Euclidean	0.0440	0.0619	closed	closed
Nbin	U (with 2 empty cells)	1.0000	120	Euclidean	0.0480	0.0450	closed	closed
Nbin	U (with 2 empty cells)	0.0500	120	Euclidean	0.0300	0.0659	conservative	liberal
Nbin	U (with 2 empty cells)	0.0200	120	Euclidean	0.0200	0.0719	conservative	liberal
Nbin	U (with 2 empty cells)	0.0100	120	Euclidean	0.0130	0.0919	conservative	liberal
Nbin	U (with 3 empty cells)	1.0000	120	Euclidean	0.0470	0.0579	closed	closed
Nbin	U (with 3 empty cells)	0.0500	120	Euclidean	0.0360	0.0709	conservative	liberal
Nbin	U (with 3 empty cells)	0.0200	120	Euclidean	0.0390	0.0879	closed	liberal
Nbin	U (with 3 empty cells)	0.0100	120	Euclidean	0.0220	0.1079	conservative	liberal

Table 29: Results (negative binomial variates under Euclidean measure when $p = 150$).

Distribution	Unbalancedness	k	N	Distance	rejections.A	rejections.B	Factor A	Factor B
Nbin	B (with no empty cells)	1.0000	60	Bray Curtis	0.0579	0.0470	closed	closed
Nbin	B (with no empty cells)	0.0500	60	Bray Curtis	0.0519	0.0549	closed	closed
Nbin	B (with no empty cells)	0.0200	60	Bray Curtis	0.0519	0.0559	closed	closed
Nbin	B (with no empty cells)	0.0100	60	Bray Curtis	0.0480	0.0539	closed	closed
Nbin	B (with 1 empty cell)	1.0000	60	Bray Curtis	0.0609	0.0450	closed	closed
Nbin	B (with 1 empty cell)	0.0500	60	Bray Curtis	0.0470	0.0549	closed	closed
Nbin	B (with 1 empty cell)	0.0200	60	Bray Curtis	0.0549	0.0440	closed	closed
Nbin	B (with 1 empty cell)	0.0100	60	Bray Curtis	0.0420	0.0490	closed	closed
Nbin	B (with 2 empty cells)	1.0000	60	Bray Curtis	0.0579	0.0559	closed	closed
Nbin	B (with 2 empty cells)	0.0500	60	Bray Curtis	0.0490	0.0579	closed	closed
Nbin	B (with 2 empty cells)	0.0200	60	Bray Curtis	0.0509	0.0549	closed	closed
Nbin	B (with 2 empty cells)	0.0100	60	Bray Curtis	0.0480	0.0480	closed	closed
Nbin	U (with no empty cells)	1.0000	60	Bray Curtis	0.0619	0.0569	closed	closed
Nbin	U (with no empty cells)	0.0500	60	Bray Curtis	0.0619	0.0480	closed	closed
Nbin	U (with no empty cells)	0.0200	60	Bray Curtis	0.0519	0.0519	closed	closed
Nbin	U (with no empty cells)	0.0100	60	Bray Curtis	0.0549	0.0549	closed	closed
Nbin	U (with no empty cells N2.>N1.)	1.0000	60	Bray Curtis	0.0649	0.0450	closed	closed
Nbin	U (with no empty cells N2.>N1.)	0.0500	60	Bray Curtis	0.0649	0.0509	closed	closed
Nbin	U (with no empty cells N2.>N1.)	0.0200	60	Bray Curtis	0.0639	0.0440	closed	closed
Nbin	U (with no empty cells N2.>N1.)	0.0100	60	Bray Curtis	0.0430	0.0519	closed	closed
Nbin	U (with no empty cells N2.<N1.)	1.0000	60	Bray Curtis	0.0529	0.0490	closed	closed
Nbin	U (with no empty cells N2.<N1.)	0.0500	60	Bray Curtis	0.0410	0.0460	closed	closed
Nbin	U (with no empty cells N2.<N1.)	0.0200	60	Bray Curtis	0.0599	0.0490	closed	closed
Nbin	U (with no empty cells N2.<N1.)	0.0100	60	Bray Curtis	0.0599	0.0599	closed	closed
Nbin	U (with 1 empty cell)	1.0000	60	Bray Curtis	0.0529	0.0589	closed	closed
Nbin	U (with 1 empty cell)	0.0500	60	Bray Curtis	0.0589	0.0539	closed	closed
Nbin	U (with 1 empty cell)	0.0200	60	Bray Curtis	0.0430	0.0569	closed	closed
Nbin	U (with 1 empty cell)	0.0100	60	Bray Curtis	0.0529	0.0490	closed	closed
Nbin	U (with 2 empty cells)	1.0000	60	Bray Curtis	0.0390	0.0539	closed	closed
Nbin	U (with 2 empty cells)	0.0500	60	Bray Curtis	0.0480	0.0420	closed	closed
Nbin	U (with 2 empty cells)	0.0200	60	Bray Curtis	0.0400	0.0649	closed	closed
Nbin	U (with 2 empty cells)	0.0100	60	Bray Curtis	0.0380	0.0559	closed	closed
Nbin	U (with 3 empty cells)	1.0000	60	Bray Curtis	0.0500	0.0569	closed	closed
Nbin	U (with 3 empty cells)	0.0500	60	Bray Curtis	0.0529	0.0519	closed	closed
Nbin	U (with 3 empty cells)	0.0200	60	Bray Curtis	0.0599	0.0619	closed	closed
Nbin	U (with 3 empty cells)	0.0100	60	Bray Curtis	0.0390	0.0549	closed	closed
Nbin	B (with no empty cells)	1.0000	120	Bray Curtis	0.0539	0.0400	closed	closed
Nbin	B (with no empty cells)	0.0500	120	Bray Curtis	0.0440	0.0500	closed	closed
Nbin	B (with no empty cells)	0.0200	120	Bray Curtis	0.0400	0.0599	closed	closed
Nbin	B (with no empty cells)	0.0100	120	Bray Curtis	0.0450	0.0480	closed	closed
Nbin	B (with 1 empty cell)	1.0000	120	Bray Curtis	0.0519	0.0500	closed	closed
Nbin	B (with 1 empty cell)	0.0500	120	Bray Curtis	0.0689	0.0390	liberal	closed
Nbin	B (with 1 empty cell)	0.0200	120	Bray Curtis	0.0529	0.0609	closed	closed
Nbin	B (with 1 empty cell)	0.0100	120	Bray Curtis	0.0539	0.0559	closed	closed
Nbin	B (with 2 empty cells)	1.0000	120	Bray Curtis	0.0539	0.0569	closed	closed
Nbin	B (with 2 empty cells)	0.0500	120	Bray Curtis	0.0440	0.0450	closed	closed
Nbin	B (with 2 empty cells)	0.0200	120	Bray Curtis	0.0400	0.0509	closed	closed
Nbin	B (with 2 empty cells)	0.0100	120	Bray Curtis	0.0450	0.0649	closed	closed
Nbin	U (with no empty cells)	1.0000	120	Bray Curtis	0.0569	0.0420	closed	closed
Nbin	U (with no empty cells)	0.0500	120	Bray Curtis	0.0460	0.0629	closed	closed
Nbin	U (with no empty cells)	0.0200	120	Bray Curtis	0.0549	0.0500	closed	closed
Nbin	U (with no empty cells)	0.0100	120	Bray Curtis	0.0509	0.0440	closed	closed
Nbin	U (with no empty cells N2.>N1.)	1.0000	120	Bray Curtis	0.0579	0.0470	closed	closed
Nbin	U (with no empty cells N2.>N1.)	0.0500	120	Bray Curtis	0.0390	0.0490	closed	closed
Nbin	U (with no empty cells N2.>N1.)	0.0200	120	Bray Curtis	0.0440	0.0529	closed	closed
Nbin	U (with no empty cells N2.>N1.)	0.0100	120	Bray Curtis	0.0400	0.0500	closed	closed
Nbin	U (with no empty cells N2.<N1.)	1.0000	120	Bray Curtis	0.0440	0.0480	closed	closed
Nbin	U (with no empty cells N2.<N1.)	0.0500	120	Bray Curtis	0.0509	0.0450	closed	closed
Nbin	U (with no empty cells N2.<N1.)	0.0200	120	Bray Curtis	0.0509	0.0509	closed	closed
Nbin	U (with no empty cells N2.<N1.)	0.0100	120	Bray Curtis	0.0739	0.0529	liberal	closed
Nbin	U (with 1 empty cell)	1.0000	120	Bray Curtis	0.0559	0.0539	closed	closed
Nbin	U (with 1 empty cell)	0.0500	120	Bray Curtis	0.0450	0.0639	closed	closed
Nbin	U (with 1 empty cell)	0.0200	120	Bray Curtis	0.0529	0.0589	closed	closed
Nbin	U (with 1 empty cell)	0.0100	120	Bray Curtis	0.0480	0.0549	closed	closed
Nbin	U (with 2 empty cells)	1.0000	120	Bray Curtis	0.0440	0.0460	closed	closed
Nbin	U (with 2 empty cells)	0.0500	120	Bray Curtis	0.0529	0.0679	closed	liberal
Nbin	U (with 2 empty cells)	0.0200	120	Bray Curtis	0.0480	0.0500	closed	closed
Nbin	U (with 2 empty cells)	0.0100	120	Bray Curtis	0.0440	0.0549	closed	closed
Nbin	U (with 3 empty cells)	1.0000	120	Bray Curtis	0.0509	0.0539	closed	closed
Nbin	U (with 3 empty cells)	0.0500	120	Bray Curtis	0.0529	0.0450	closed	closed
Nbin	U (with 3 empty cells)	0.0200	120	Bray Curtis	0.0420	0.0539	closed	closed
Nbin	U (with 3 empty cells)	0.0100	120	Bray Curtis	0.0390	0.0490	closed	closed

Table 30: Results (negative binomial variates under Bray Curtis measure when $p = 150$).